PCT/EP 99/06081

BUNDESREPUBLIK DEUTSCHLAND

09/763149

EP99/6081

PRIORITY DOCUMENT
SUBMITTED OR TRANSMITTED IN COMPLIANCE WITH RULE 17.1(a) OR (b)

REC'D 28 SEP 1999
WIPO PCT

Bescheinigung

4

Herr Christoph B u s k i e s in Wiesbaden/Deutschland hat eine Patentanmeldung unter der Bezeichnung

"Verfahren und Vorrichtungen zur koartikulationsgerechten Konkatenation von Audiosegmenten sowie Vorrichtungen zur Bereitstellung koartikulationsgerecht konkatenierter Audiodaten"

am 19. August 1998 beim Deutschen Patent- und Markenamt eingereicht.

Die angehefteten Stücke sind eine richtige und genaue Wiedergabe der ursprünglichen Unterlagen dieser Patentanmeldung.

Die Anmeldung hat im Deutschen Patent- und Markenamt vorläufig das Symbol G 10 L 13/00 der Internationalen Patentklassifikation erhalten.

München, den 25. August 1999
Deutsches Patent- und Markenamt
Der Präsident
Im Auftrag

Aktenzeichen: 198 37 661.8

Keller

A 9161
06.90
11/98

06.09.99

Verfahren und Vorrichtungen zur koartikulationsgerechten Konkatenation von Audiosegmenten sowie Vorrichtungen zur Bereitstellung koartikulationsgerecht konkatenierter Audiodaten

Die Erfindung betrifft ein Verfahren und eine Vorrichtung zur Konkatenation von Audiosegmenten zur Erzeugung synthetisierter akustischer Daten, insbesondere synthetisierter Sprache, gemäß den Ansprüchen 1 und 16. Des weiteren betrifft die Erfindung synthetisierte Sprachsignale gemäß Anspruch 32, die durch die erfindungsgemäße koartikulationsgerechte Konkatenation von Sprachsegmenten erzeugt wurden, sowie einen Datenträger gemäß Anspruch 45, der ein Computerprogramm zur erfindungsgemäßen Herstellung von synthetisierten akustischen Daten, insbesondere synthetisierter Sprache, enthält.

Zusätzlich betrifft die Erfindung einen Datenspeicher gemäß Anspruch 58, der Audiosegmente enthält, die zur erfindungsgemäßen koartikulationsgerechten Konkatenation geeignet sind, und einen Tonträger nach Anspruch 67, der erfindungsgemäß synthetisierte akustische Daten enthält sowie einen Tonträger nach Anspruch 69, der synthetisierte Sprachdaten gemäß Anspruch 32 enthält.

Es ist zu betonen, daß sowohl der im folgenden dargestellte Stand der Technik als auch die vorliegenden Erfindung den gesamten Bereich der Synthese von akustischen Daten durch Konkatenation einzelner, auf beliebige Art und Weise erhaltene Audiosegmente betrifft. Aber um die Diskussion des Standes der Technik sowie die Beschreibung der vorliegenden Erfindung zu vereinfachen, beziehen sich die folgenden Ausführungen speziell auf synthetisierte Sprachdaten durch Konkatenation einzelner Sprachsegmente.

In den letzten Jahren hat sich im Bereich der Sprachsynthese der datenbasierte Ansatz gegenüber dem regelbasierten Ansatz durchgesetzt und ist in verschiedenen Verfahren und Systemen zur Sprachsynthese zu finden. Obwohl der regelbasierte Ansatz prinzipiell eine bessere Sprachsynthese ermöglicht, ist es für dessen Umsetzung notwendig, das gesamte zur Spracherzeugung notwendige Wissen explizit zu formulieren, d.h. die zu synthetisierende Sprache formal zu modellieren. Da die bekannten Sprachmodellierungen Vereinfachung der zu synthetisierenden Sprache aufweisen, ist die Sprachqualität der so erzeugten Sprache nicht ausreichend.

Daher wird in zunehmenden Maße eine datenbasierte Sprachsynthese durchgeführt, bei der aus einer einzelne Sprachsegmente aufweisenden Datenbasis entsprechende Segmente ausgewählt und miteinander verknüpft (konkateniert) werden. Die Sprachqualität hängt hierbei in erster Linie von der Zahl und Art der verfügbaren Sprachsegmente ab, denn es kann nur Sprache synthetisiert werden, die durch Sprachsegmente in der Datenbasis wiedergeben ist. Um die Zahl der vorzusehenden Sprachsegmente zu minimieren und dennoch eine synthetisierte Sprache hoher Qualität zu erzeugen, sind verschieden Verfahren bekannt, die eine Verknüpfung (Konkatenation) der Sprachsegmente nach komplexen Regeln durchführen.

Unter Verwendung solcher Verfahren bzw. entsprechender Vorrichtungen kann ein Inventar, d.h. eine die Sprachsegmente umfassende Datenbasis, verwendet werden, das vollständig und handhabbar ist. Ein Inventar ist vollständig, wenn damit jede Lautfolge der zu synthetisierenden Sprache erzeugt werden kann, und ist handhabbar, wenn die Zahl und Art der Daten des Inventars mit den technisch verfügbaren Mitteln in einer gewünschten Weise verarbeitet werden kann. Darüber hinaus muß ein solches Verfahren gewährleisten, daß die Konkatenation der einzelnen Inventarelemente eine synthetisierte Sprache erzeugt, die sich von einer natürlich gesprochenen Sprache möglichst wenig unterscheidet. Hierfür muß eine synthetisierte Sprache flüssig sein und die gleichen artikulatorischen Effekte einer natürlichen

Sprache aufweisen. Hier kommen den sogenannten koartikulatorischen Effekten, d.h. der gegenseitigen Beeinflussung von Sprachlauten, eine besondere Bedeutung zu. Daher sollten die Inventarelemente so beschaffen sein, das sie die Koartikulation einzelner aufeinanderfolgender Sprachlaute berücksichtigen. Des weiteren sollte ein Verfahren zu Konkatenation der Inventarelemente, die Elemente unter Berücksichtigung der Koartikulation einzelner aufeinanderfolgender Sprachlaute sowie der übergeordneten Koartikulation mehrerer aufeinanderfolgender Sprachlaute, auch über Wort- und Satzgrenzen hinweg, verketten.

Vor der Darstellung des Standes der Technik werden im folgenden einige zum besseren Verständnis notwendige Begriffe aus dem Bereich der Sprachsynthese erläutert:

- Ein Phonem ist die kleinste formal beschreibbare Lauteinheit, wobei i. allg. die formale Beschreibung durch Lautschriftzeichen erfolgt.
- Ein Phon ist die kleinste Lauteinheit, die in Form eines Audiosegmentes speicherbar ist, und stellt die akustische Realisierung eines Phonems dar. Die Phone werden in statische und dynamische Phone unterteilt.
- Zu den statischen Phonen zählen Vokale, Diphtonge, Nasale, Laterale, Vibranten und Frikative.
- Zu den dynamischen Phonen zählen Plosive, Affrikate, Glottalstops und geschlagene Laute.
- Die Koartikulation bezeichnet das Phänomen, daß ein Phon durch vorgelagerte und nachgelagerte Phone beeinflußt wird, wobei die Koartikulation zwischen unmittelbar benachbarten Phonen auftritt, aber sich auch über eine Folge mehrerer Phone erstrecken kann (Beispielsweise bei einer Lippenrundung).

Daher kann ein Phon in drei Bereiche unterteilt werden (siehe auch Figur 1b):
- Der Anfangs-Koartikulationsbereich umfaßt den Bereich vom Beginn des Phons bis zum Ende der Koartikulation aufgrund eines vorgelagerten Phons.

- Der Solo-Artikulationsbereich, ist der Bereich des Phons, der nicht durch ein vor- oder nachgelagertes Phon beeinflußt ist.
- Der End-Koartikulationsbereich umfaßt den Bereich vom Beginn der Koartikulation aufgrund eines nachgelagerten Phons bis zum Ende des Phons.

- Ein Polyphon ist eine Folge von Phonen.
- Die Elemente eines Inventars sind in kodierter Form gespeicherte Audiosegmente, die Phone, Teile von Phonen oder Polyphone wiedergeben. Zur besseren Verständnis des möglichen Aufbau eines Elementarelementes sei hier auf die Figur 2a, die ein herkömmliches Audiosegment zeigt, und die Figuren 2b-21 verwiesen, in denen erfindungsgemäße Audiosegmente gezeigt sind. Ergänzend ist zu erwähnen, daß Audiosegmente auch aus kleineren oder größeren Audiosegmenten gebildet werden können, die in dem Inventar oder einer Datenbank enthalten sind. Des weiteren können Audiosegmente auch in einer transformierten Form (z.B. einer fouriertransformierten Form) in dem Inventar oder einer Datenbank vorliegen.
- Unter Konkatenation versteht man das Aneinanderfügen zweier Inventarelemente.
- Der Konkatenationsmoment ist der Zeitpunkt, zu dem zwei Audiodaten aneinandergefügt werden.

Die Konkatenation kann auf verschiedene Arten erfolgen, z.B. mit einem Crossfade oder einem Hardfade (siehe auch Figuren 3a-3d):
- Bei einem Crossfade werden ein zeitlich hinterer Bereich eines ersten Inventarelementes sowie ein zeitlich vorderer Bereich eines zweiten Inventarelementes geeignet gewichtet, und danach werden diese beiden Bereiche überlappend so addiert, daß maximal der zeitliche kürzer der beiden Bereichen von dem zeitlich längeren der beiden Bereiche vollständig überlappt wird.
- Bei einem Hardfade wird ein zeitlich hinterer Bereich eines ersten Inventarelementes und ein zeitlich vorderer Bereich eines zweiten Inventarelementes geeignet gewichtet, wobei diese beiden Inventarelemente so aneinandergefügt werden, daß sich

der hintere Bereich des ersten Inventarelementes und der vordere Bereich des zweiten Inventarelementes nicht überlappen.

Der Koartikulationsbereich macht sich vor allem dadurch bemerkbar, daß eine Konkatenation darin mit Unstetigkeiten (z.B. Spektralsprüngen) verbunden ist. Deswegen wird der Konkatenationsmoment vorzugsweise in der Umgebung der Grenze des Solo-Artikulationsbereiches zum Koartikulationsbereich gewählt.

Im allgemeinen werden Inventarelemente durch die Aufnahme von real gesprochener Sprache erzeugt. In Abhängigkeit des Trainingsgrades des inventaraufbauenden Sprechers, d.h. seiner Fähigkeit die aufzunehmende Sprache zu kontrollieren (z.B. die Tonhöhe der Sprache zu kontrollieren oder exakt auf einer Tonhöhe zu sprechen), ist es möglich, gleiche oder ähnliche Inventarelemente zu erzeugen, die verschobene Grenzen zwischen den Solo-Artikulationsbereichen und Koartikulationsbereichen haben. Dadurch ergeben sich wesentlich mehr Möglichkeiten, die Konkatenationspunkte an verschiedenen Stellen zu plazieren. In der Folge kann die Qualität einer zu synthetisierenden Sprache deutlich verbessert werden.

Ergänzend sei zu erwähnen, daß streng genommen ein Hardfade einen Grenzfall eines Crossfades darstellt, bei dem eine Überlappung eines zeitlich hinteren Bereiches eines ersten Inventarelementes und eines zeitlich vorderen Bereiches eines zweiten Inventarelementes eine Länge Null hat. Dies erlaubt es in bestimmten, z.B. äußerst zeitkritischen Anwendungen einen Crossfade durch einen Hardfade zu ersetzen, wobei eine solche Vorgehensweise genau abzuwägen ist, da diese zu deutlichen Qualitätseinbußen bei der Konkatenation von Inventarelementen führt, die eigentlich durch einen Crossfade zu konkatenieren sind.

- Unter <u>Prosodie</u> versteht man die Veränderungen der Sprachfrequenz und des Sprachrhythmus, die bei gesprochenen Worten bzw. Sätzen auftreten. Die Berücksichtigung solcher prosodischer Informationen ist bei der Sprachsynthese notwendig, um eine natürliche Wort- bzw. Satzmelodie zu erzeugen.

Aus WO 95/30193 ist ein Verfahren und eine Vorrichtung zur Umwandlung von Text in hörbare Sprachsignale unter Verwendung eines neuronalen Netzwerkes bekannt. Hierfür wird der in Sprache umzuwandelnde Text mit einer Konvertiereinheit in eine Folge von Phonemen umgewandelt, wobei zusätzlich Informationen über die syntaktischen Grenzen des Textes und die Betonung der einzelnen syntaktischen Komponenten des Textes erzeugt werden. Diese werden zusammen mit den Phonemen an eine Einrichtung weitergeleitet, die regelbasiert die Dauer der Aussprache der einzelnen Phoneme bestimmt. Ein Prozessor erzeugt aus jedem einzelnen Phonem in Verbindung mit den entsprechenden syntaktischen und zeitlichen Information eine geeignet Eingabe für das neuronale Netzwerk, wobei diese Eingabe für das neuronale Netz auch die entsprechenden prosodischen Informationen für die gesamte Phonemfolge umfaßt. Das neuronale Netz wählt aus den verfügbaren Audiosegmenten nun die aus, die die eingegebenen Phoneme am besten wiedergeben, und verkettet diese Audiosegmente entsprechend. Bei dieser Verkettung werden die einzelnen Audiosegmente in ihrer Dauer, Gesamtamplitude und Frequenz an vor- und nachgelagerte Audiosegmente unter Berücksichtigung der prosodischen Informationen der zu synthetisierenden Sprache angepaßt und zeitlich aufeinanderfolgend miteinander verbunden. Eine Veränderung einzelner Bereiche der Audiosegmente ist hier nicht beschrieben.

Zur Erzeugung der für dieses Verfahren erforderlichen Audiosegmente ist das neuronale Netzwerk zuerst zu trainieren, indem natürlich gesprochene Sprache in Phone oder Phonfolgen unterteilt wird und diesen Phonen oder Phonfolgen entsprechende Phonem oder Phonemfolgen in Form von Audiosegmenten zugeordnet werden. Da dieses Verfahren nur eine Veränderung von einzelnen Audiosegmenten, aber keine Veränderung einzelner Bereiche eines Audiosegmentes vorsieht, muß das neuronale Netzwerk mit möglichst vielen verschiedenen Phonen oder Phonfolgen trainiert werden, um beliebige Texte in synthetisierte natürlich klingende Sprache umzuwandeln. Dies kann sich je nach Anwendungsfall sehr aufwendig gestalten. Auf der anderen Seite kann ein unzu-

reichender Trainingsprozeß des neuronalen Netzes die Qualität der zu synthetisierenden Sprache negativ beeinflussen. Des weiteren ist es bei dem hier beschriebene Verfahren nicht möglich, den Konkatenationsmoment der einzelnen Audiosegmente in Abhängigkeit vorgelagerter oder nachgelagerter Audiosegmente zu bestimmen, um so eine koartikulationsgerechte Konkatenation durchzuführen.

In US-5,524,172 ist eine Vorrichtung zur Erzeugung synthetisierter Sprache beschrieben, die das sogenannte Diphonverfahren nutzt. Hier wird ein Text, der in synthetisierte Sprache umgewandelt werden soll, in Phonemfolgen unterteilt, wobei jeder Phonemfolge entsprechende prosodische Informationen zugeordnet werden. Aus einer Datenbank, die Audiosegmente in Form von Diphonen enthält, werden für jedes Phonem der Folge zwei das Phonem wiedergebende Diphone ausgewählt und unter Berücksichtigung der entsprechenden prosodischen Informationen konkateniert. Bei der Konkatenation werden die beiden Diphone jeweils mit Hilfe eines geeigneten Filters gewichtet und die Dauer und Tonhöhe beider Diphone so verändert, daß bei der Verkettung der Diphone eine synthetisierte Phonfolge erzeugt wird, deren Dauer und Tonhöhe der Dauer und Tonhöhe der gewünschten Phonemfolge entspricht. Bei der Konkatenation werden die einzelnen Diphone so addiert, daß sich ein zeitlich hinterer Bereich eines ersten Diphones und ein zeitlich vorderer Bereich eines zweiten Diphones überlappen, wobei der Konkatenationsmoment generell im Bereich stationären Bereiche der einzelnen Diphone liegt (siehe Figur 2a). Da eine Variation des Konkatenationsmomentes unter Berücksichtigung der Koartikulation aufeinanderfolgender Audiosegmente (Diphone) hier nicht vorgesehen ist, kann die Qualität (Natürlichkeit und Verständlichkeit) einer so synthetisierten Sprache negativ beeinflußt werden.

Eine Weiterentwicklung des zuvor diskutierten Verfahrens ist in EP-0,813,184 A1 zu finden. Auch hier wird ein in synthetisierte Sprache umzuwandelnder Text in einzelne Phoneme oder Phonemfolgen unterteilt und aus einer Datenbank entsprechende Audioseg-

mente ausgewählt und konkateniert. Um eine Verbesserung der synthetisierten Sprache zu erzielen, sind bei diesem Verfahren zwei Ansätze, die sich vom bisher diskutierten Stand der Technik unterscheiden, umgesetzt worden. Unter Verwendung eines Glättungsfilters, der die tieferfrequenten harmonischen Frequenzanteile eines vorgelagerten und eines nachgelagerten Audiosegmentes berücksichtigt, soll der Übergang von dem vorgelagerten Audiosegment zu dem nachgelagerten Audiosegment optimiert werden, indem ein zeitlich hinterer Bereich des vorgelagerten Audiosegmentes und ein zeitlich vorderer Bereich des nachgelagerten Audiosegmentes im Frequenzbereich aufeinander abgestimmt werden. Des weiteren stellt die Datenbank Audiosegmente zur Verfügung, die sich leicht unterscheiden, aber zur Synthetisierung desselben Phonems geeignet sind. Auf diese Weise soll die natürliche Variation der Sprache nachgebildet werden, um eine höhere Qualität der synthetisierten Sprache zu erreichen. Sowohl die Verwendung des Glättungsfilter als auch die Auswahl aus einer Menge unterschiedlicher Audiosegmente zur Realisierung eines Phonems erfordert bei einer Umsetzung dieses Verfahrenes eine hohe Rechenleistung der verwendeten Systemkomponenten. Außerdem steigt der Umfang der Datenbank aufgrund der erhöhten Zahl der vorgesehenen Audiosegmente. Des weiteren ist auch bei diesem Verfahren eine koartikulationsabhängige Wahl des Konkatenationsmomentes einzelner Audiosegmente nicht vorgesehen, wodurch die Qualität der synthetisierten Sprache reduziert werden kann.

Zusammenfassend ist zu sagen, daß es der Stand der Technik zwar erlaubt, beliebige Phonemfolgen zu synthetisieren, aber die so synthetisierten Phonemfolgen haben keine authentische Sprachqualität. Eine synthetisierte Phonemfolge hat eine authetische Sprachqualität, wenn sie von der gleichen Phonemfolge, die von einem realen Sprecher gesprochen wurde, durch einen Hörer nicht unterschieden werden kann.

Es sind auch Verfahren bekannt, die ein Inventar benutzen, das vollständige Worte und/oder Sätze in authentischer Sprachqualität als Inventarelemente enthält. Diese Elemente werden zur

Sprachsynthese in einer gewünschten Reihenfolge hintereinander gesetzt, wobei die Möglichkeiten unterschiedliche Sprachsequenzen in hohem Maße von dem Umfang eines solchen Inventars limitiert werden. Die Synthese beliebiger Phonemfolgen ist mit diesen Verfahren nicht möglich.

Daher ist es eine Aufgabe der vorliegenden Erfindung ein Verfahren und eine entsprechende Vorrichtung zur Verfügung zu stellen, die die Probleme des Standes der Technik beseitigen und die Erzeugung synthetisierter akustischer Daten, insbesondere synthetisierter Sprachdaten, ermöglichen, die sich für einen Hörer nicht von entsprechenden natürlichen akustischen Daten, insbesondere natürlich gesprochener Sprache, unterscheiden. Die mit der Erfindung synthetisierten akustischen Daten, insbesondere synthetisierte Sprachdaten sollen eine authentische akustische Qualität, insbesondere eine authentische Sprachqualität aufweisen.

Zu Lösung dieser Aufgabe sieht die Erfindung ein Verfahren gemäß Anspruch 1 und eine Vorrichtung gemäß Anspruch 16 vor. Dabei wird zur Erzeugung synthetisierter akustischer Daten, die aus einer Folge von Lauteinheiten bestehen, durch Konkatenation von Audiosegmenten der Moment der Konkatenation zweier Audiosegmente in Abhängigkeit von Eigenschaften der zu verknüpfenden Audiosegmente, insbesondere der die beiden Audiosegmente betreffenden Koartikulationseffekte bestimmt. Auf diese Weise wird eine Sprachqualität erreicht, die mit dem Stand der Technik nicht erzielbar ist. Dabei ist die erforderliche Rechenleistung nicht höher als beim Stand der Technik.

Eine weitere Aufgabe der Erfindung ist es, bei der Synthese akustischer Daten die Variationen nachzubilden, die bei entsprechenden natürlichen akustischen Daten zu finden sind. Daher sieht das erfindungsgemäße Verfahren Schritte zur unterschiedlichen Auswahl der Audiosegmente nach den Ansprüchen 2 oder 5 sowie unterschiedliche Arten der Konkatenation nach den Ansprüchen 3 oder 4 vor. Ebenso stellt die erfindungsgemäße Vorrichtung unterschiedliche Audiosegmente nach den Ansprüchen 16 oder

20 zur Verfügung und ermöglicht unterschiedliche Konkatenationsarten nach den Ansprüchen 18 oder 19, die in Abhängigkeit von Eigenschaften der zu verkettenden Audiosegmente gewählt werden. So wird ein höheres Maß an Natürlichkeit der synthetisierten akustischen Daten erzielt. Vorzugsweise werden die Konkatenationen nach den Ansprüchen 3 oder 4 unter Verwendung eines Crossfades oder eines Hardfades durchgeführt bzw. die Einrichtungen nach den Ansprüchen 18 oder 19 sind zu Durchführung eines Crossfades oder Hardfades zu Konkatenation der Audiosegmente geeignet.

Eine weitere Aufgabe der Erfindung ist es, die Konkatenation der einzelnen Audiosegmente zu optimieren, um die Erzeugung der synthetisierten akustischen Daten einfacher und schneller durchzuführen. Zur Lösung dieser Aufgabe umfaßt das erfindungsgemäße Verfahren Schritte nach den Ansprüchen 6, 7 oder 8, die es ermöglichen die Zahl der zur Datensynthetisierung notwendigen Audiosegmente zu reduzieren. In ähnlicher Weise stellt die erfindungsgemäße Vorrichtungen Einrichtungen nach den Ansprüchen 22, 23 oder 24 zur Verfügung, die Audiosegmente vorsieht oder erzeugt, die eine einfachere und schnellere Erzeugung synthetisierter akustischer Daten erlauben. Auf diese Weise kann auch mit Vorrichtungen, die eine geringere Rechenleistung haben (z.B. Anrufbeantworter oder Autoleitsysteme), ein synthetisierter Sprache hoher Qualität erzeugt werden. Des weiteren sinkt der zur Speicherung des Inventars notwendige Speicherbedarf.

Eine andere Aufgabe der Erfindung ist es, bei der Erzeugung der synthetisierten akustischen Daten akustische Phänomene nachzubilden, die sich aufgrund einer gegenseitigen Beeinflussung einzelner Segmente entsprechender natürlicher akustischer Daten ergeben. Daher sieht das erfindungsgemäße Verfahren Schritte nach den Ansprüchen 9 oder 10 vor bzw. umfaßt die erfindungsgemäße Vorrichtung Einrichtungen nach den Ansprüchen 25 oder 26, die zur Nachbildung dieser Phänomene geeignet sind. Insbesondere ist hier vorgesehen, einzelne Audiosegmente bzw. einzelne Bereiche der Audiosegmente in ihrer Frequenz, Dauer und Ampli-

tude(n) zu variieren. Werden mit der Erfindung synthetisierte Sprachdaten erzeugt, so werden zur Lösung dieser Aufgabe vorzugsweise prosodische Informationen und/oder übergeordnete Koartikulationseffekte berücksichtigt.

Des weiteren soll die Erfindung ein Verfahren bzw. eine Vorrichtung zur Verfügung stellen, die den Signalverlauf von synthetisierten akustischen Daten verbessern. Zur Lösung dieser Aufgabe sieht die Erfindung ein Verfahren nach Anspruch 11 bzw. eine Vorrichtung nach Anspruch 27 vor, die es ermöglichen, den Konkatenationsmoment an Nullstellen der einzelnen zu verknüpfenden Audiosegmente zu legen.

Eine weitere andere Aufgabe der Erfindung ist es, die Auswahl der Audiosegmente zur Erzeugung der synthetisierten akustischen Daten zu verbessern sowie deren Konkatenation effizienter zu gestalten. Diese Aufgabe wird durch die Nutzung heuristischen Wissens gelöst, das die Auswahl, Variation und Konkatenation der Audiosegmente betrifft, wobei die Lösung dieser Aufgabe durch einen erfindungsgemäßen Verfahrensschritt nach Anspruch 12 bzw. durch ein Merkmal der erfindungsgemäßen Vorrichtung nach Anspruch 28 ermöglicht wird.

Außerdem soll Erfindung die Nutzung der erzeugten synthetisierten akustischen Daten möglich machen. Daher werden unter Verwendung des erfindungsgemäßen Verfahrens nach den Ansprüchen 13 oder 14 synthetisierte akustische Daten zur Verfügung gestellt, die zur Weiterverarbeitung in nachgelagerten Schritten geeignet sind, wobei diese Daten vorzugsweise in akustische Signale umwandelbar oder auf einem Datenträger speicherbar sind. Ebenso umfaßt die erfindungsgemäße Vorrichtung Einrichtungen nach den Ansprüchen 29 oder 30, die erzeugte synthetisierte akustische Daten zur Weiterverarbeitung vorbereiten, vorzugsweise zur akustischen Wiedergabe oder datentechnischen Speicherung.

Ein weiteres Ziel dieser Erfindung ist es, synthetisierte Sprachdaten zu erzeugen, die sich von entsprechenden natürlichen Sprachdaten nicht unterscheiden. Diese Aufgabe wird durch

das erfindungsgemäße Verfahren dadurch gelöst, daß nach Anspruch 15 bei dessen Durchführung Audiosegmente genutzt werden, die Phone oder Polyphone wiedergeben, und durch die erfindungsgemäße Vorrichtung dadurch gelöst, daß diese Einrichtungen nach Anspruch 31 umfaßt, die Audiosegmente in Form von Phonen oder Polyphonen vorsehen und die zur Konkatenation dieser Audiosegmente geeignet sind.

Eine andere Aufgabe der Erfindung ist es, synthetisierte Sprachsignale zu Verfügung zu stellen, die sich von bekannten synthetisierten Sprachsignalen dadurch unterscheiden, daß sie sich in ihrer Natürlichkeit und Verständlichkeit nicht von realer Sprache unterscheiden. Hierfür sieht Erfindung Sprachsignale gemäß Anspruch 32 vor, die aus einer Folge von Phonen bestehen und durch Konkatenation von Audiosegmenten erzeugt werden, wobei der Moment der Konkatenation zweier Audiosegmente in Abhängigkeit von Eigenschaften der zu verknüpfenden Audiosegmente, insbesondere der die beiden Audiosegmente betreffenden Koartikulationseffekte, bestimmt wird.

Eine weitere Aufgabe der Erfindung ist es, synthetisierte Sprachsignale bereitzustellen, die die Variationen und gegenseitige Beeinflussungen widergeben, die bei entsprechenden natürlichen Sprachsignalen zu finden sind. Daher stellt die Erfindung auch synthetisierte Sprachsignale nach den Ansprüchen 33 bis 37 zur Verfügung. Ein andere weitere Aufgabe ist es, Sprachsignale schneller zur Verfügung zu stellen bzw. Sprachsignale, zur Verfügung zu stellen, die eine verringerte Anzahl von Konkatenationsmomenten haben, um eine verbesserte Natürlichkeit und Verständlichkeit dieser Sprachsignale zu erzielen. Diese Aufgabe wird durch Sprachsignale gelöst, die Merkmale nach den Ansprüchen 37, 38 oder 39 aufweisen.

Zusätzlich ist es eine Aufgabe der Erfindung, Sprachsignale vorzusehen, die einen natürlichen Sprachfluß, Sprachmelodie und Sprachrhythmus haben. Daher stellt die Erfindung auch Sprachsignale zur Verfügung, die Merkmale der Ansprüche 40 und/oder 41 aufweisen. Vorzugsweise umfassen die synthetisierten Sprachsi-

gnale solche Audiosegmente in Form von Phonen oder Phonfolgen, die jeweils vor und/oder nach der Konkatenation in ihrer Gesamtheit oder in einzelnen Bereichen in ihrer Frequenz, Dauer und Amplitude variiert werden.

Des weiteren sollen erfindungsgemäße Sprachsignale einen verbesserten Signalverlauf aufweisen. Zur Lösung dieser Aufgabe stellt die Erfindung Sprachsignale nach Anspruch 42 zur Verfügunng, die Konkatenationsmomente aufweisen, die an Nullstellen der zu verknüpfenden Audiosegmente liegt.

Des weiteren sollen die erfindungsgemäßen Sprachsignale eine allgemeine Nutzung und/oder Weiterverarbeitung durch bekannte Verfahren oder Vorrichtungen, z.B. einem CD-Abspielgerät, erlauben. Deshalb sieht die Erfindung Sprachsignale nach den Ansprüchen 43 und/oder 44 vor, die vorzugsweise in akustische Signale umwandelbar oder auf einem Datenträger speicherbar sind.

Eine andere Aufgabe der Erfindung ist es synthetisierte akustische Daten, insbesondere synthetisierte Sprachdaten, zu erzeugen, die sich von entsprechenden natürlichen akustischen Daten nicht unterscheiden, wobei die Erzeugung dieser Daten unter Verwendung bekannter Vorrichtungen, z.B. einem Personal Computer oder einem computergesteuerten Musikinstrument, durchgeführt wird. Hierfür sieht die Erfindung einen Datenträger nach Anspruch 45 vor, der ein Computerprogramm enthält, das Audiosegmente auswählt und durch Konkatenation zu synthetisierten akustischen Daten verkettet, wobei der Moment der Konkatenation zweier Audiosegmente in Abhängigkeit von Eigenschaften der zu verknüpfenden Audiosegmente, insbesondere der die beiden Audiosegmente betreffenden Koartikulationseffekte, bestimmt wird.

Eine weitere Aufgabe der Erfindung ist es, bei der Synthese akustischer Daten unter Verwendung des Datenträgers nach Anspruch 45 die Variationen nachzubilden, die bei entsprechenden natürlichen akustischen Daten zu finden sind. Daher stellt die Erfindung einen Datenträger zur Verfügung, der ein Computerpro-

gramm enthält, das nach Ansprüchen 46 und/oder 49 in Abhängigkeit der zu erzeugenden Daten Audiosegmente unterschiedlich auswählt bzw. das nach den Ansprüchen 47 und/oder 48 einzelne Audiosegmente in Abhängigkeit von Eigenschaften der zu verkettenden Audiosegmente unterschiedlich konkateniert.

Eine andere Aufgabe der Erfindung ist es, ein Computerprogramm vorzusehen, das die Konkatenation einzelner Audiosegmente optimiert, um die Erzeugung der synthetisierten akustischen Daten einfacher und schneller durchzuführen. Diese Aufgabe wird durch einen erfindungsgemäßen Datenträger gelöst, der ein Computerprogramm enthält, das die Merkmale der Ansprüche 50 und/oder 51 aufweist.

Eine weitere andere Aufgabe der Erfindung ist es, mit Hilfe eines Computerprogrammes bei der Erzeugung der synthetisierten akustischen Daten die akustischen Phänomene nachzubilden, die sich aufgrund einer gegenseitigen Beeinflussung einzelner Segmente entsprechender natürlicher akustischer Daten ergeben. Daher sieht die Erfindung einen Datenträger vor, der ein Computerprogramm mit den Merkmale der Ansprüche 51 und/oder 52 enthält. Vorzugsweise soll das Computerprogramm die Variation der Frequenzen, Dauer und Amplituden einzelner Audiosegmente bzw. einzelner Bereiche der Audiosegmente ermöglichen. Dient das Computerprogramm zur Erzeugung synthetisierter Sprachdaten, so werden zur Lösung dieser Aufgabe vorzugsweise prosodische Informationen und/oder übergeordnete Koartikulationseffekte berücksichtigt.

Außerdem soll die Erfindung ein Computerprogramm vorsehen, das eine Verbesserung des Signalverlaufes von synthetisierten akustischen Daten ermöglicht. Diese Aufgabe wird durch einen erfindungsgemäßen Datenträger gelöst, der ein Computerprogramm mit den Merkmalen des Anspruches 53 enthält.

Eine zusätzliche Aufgabe der Erfindung ist es, ein Computerprogramm zur Verfügung zu stellen, das es erlaubt, die synthetisierte akustische Daten, insbesondere synthetisierte Sprach-

signale, zu erzeugen, wobei die Auswahl, Variation und Konkatenation einzelner Audiosegmente nicht auf der Basis einer formalen Modellierung durchgeführt wird. Zur Lösung stellt die Erfindung einen Datenträger nach Anspruch 54 bereit, der unter Verwendung eines darauf enthaltenen Computerprogrammes heuristisches Wissen implementiert, das die Auswahl, Variation und/oder Konkatenation einzelner Audiosegmente betrifft. Auf diese Weise ist es möglich mit zunehmender Dauer der Verwendung des Computerprogrammes eine immer höhere Qualität, d.h. z.B. Natürlichkeit, der synthetisierten akustischen Daten zu erreichen.

Außerdem soll ein erfindungsgemäßes Computerprogramm die Nutzung und/oder Weiterverarbeitung der erzeugten synthetisierten akustischen Daten mit bekannten Vorrichtungen, z.B. einem Tonbandgerät, möglich machen. Zur Lösung dieser Aufgabe umfaßt die Erfindung einen Datenträger, der ein Computerprogramm nach den Ansprüchen 55 und/oder 56 enthält, wobei das Computerprogramm vorzugsweise Daten erzeugt, die in akustische Signale umwandelbar oder auf einem Datenträger speicherbar sind.

Darüber hinaus ist es eine Aufgabe der Erfindung mit Hilfe eines Computerprogrammes synthetisierte Sprachdaten zu erzeugen, die sich von entsprechenden natürlichen Sprachdaten nicht unterscheiden. Hierzu stellte die Erfindung einen Datenträger nach Anspruch 57 bereit, der ein Computerprogramm enthält, das Audiosegmente, die Phone oder Polyphone wiedergeben, zu synthetischen Sprachsignalen konkateniert.

Eine andere Aufgabe der Erfindung ist es, ein Audiosegmente umfassendes Inventar und insbesondere ein Sprachsegmente umfassendes Inventar vorzusehen, mit dem synthetisierte akustische Daten, insbesondere synthetisierte Sprachdaten, erzeugt werden können, die sich von entsprechenden natürlichen akustischen Daten nicht unterscheiden. Zur Lösung dieser Aufgabe sieht die Erfindung einen Datenspeicher nach Anspruch 58 vor, der Audiosegmente enthält, die geeignet sind, um erfindungsgemäß zu synthetisierten akustischen Daten konkateniert zu werden.

Vorzugsweise enthält ein solcher Datenträger Audiosegmente, die nach Anspruch 59 Phone und/oder nach Anspruch 60 Polyphone wiedergeben. Des weiteren ist zu bevorzugen, daß der Datenträger Audiosegmente enthält, die die Merkmale der Anspruches 61 und/oder 62 aufweisen.

Eine weitere andere Aufgabe ist es, ein Inventar zur Verfügung zu stellen, das die Erzeugung synthetisierter akustischer Daten und insbesondere die Erzeugung synthetisierter Sprachdaten erlaubt, die unter Berücksichtigung von akustischer Effekte durchgeführt wird, die auf eine gegenseitige Beeinflussung der verwendeten Audiosegmente zurückzuführen sind. Daher umfaßt der Datenträger zusätzliche die Audiosegmente betreffende Informationen nach den Ansprüchen 63 und/oder 64. Vorzugsweise betreffen diese Informationen die Variation der Frequenzen, Dauer und Amplituden einzelner Audiosegmente oder einzelner Bereiche von Audiosegmenten. Werden Audiosegmente verwendet, die Phone und/oder Polyphone wiedergeben, so sind diese Informationen vorzugsweise prosodische Informationen und/oder übergeordnete Koartikulationsphänomene betreffenden Informationen. Außerdem soll ein Datenspeicher zur Verfügung gestellt werden, dessen Inventar eine Verbesserung des Signalverlaufes synthetisierter akustischer Daten ermöglicht. Diese Aufgabe wird durch Verwendung eines Datenträgers nach Anspruch 65 gelöst. Des weiteren ist hierfür zu bevorzugen, daß diese Information zusätzlich Merkmale des Anspruches 66 aufweisen, um durch die Nutzung heuristischen Wissens, das die Auswahl, Variation und/oder Konkatenation einzelner Audiosegmente betrifft, die Qualität der erzeugten synthetisierten akustischen Daten und insbesondere der erzeugten synthetisierten Sprachdaten zu verbessern.

Schließlich ist es eine Aufgabe der Erfindung, erfindungsgemäße synthetisierte akustische Daten, insbesondere synthetisierte Sprachdaten, zur Verfügung zu stellen, die mit herkömmlichen bekannten Vorrichtungen, beispielsweise einem Tonbandgerät oder einer PC-Audiokarte, genutzt werden können. Diese Aufgabe wird durch die Bereitstellung eine Tonträgers nach den Ansprüchen 67, 68 bzw. 69 gelöst.

Weitere Eigenschaften, Merkmale, Vorteile oder Abwandlungen der Erfindung werden anhand der nachfolgenden Beschreibung erläutert. Dabei zeigt:

Figur 1a: Schematische Darstellung einer erfindungsgemäßen Vorrichtung zur Erzeugung synthetisierter akustischer Daten;

Figur 1b: Struktur eines Phons;

Figur 2a: Struktur eines herkömmlichen Audiosegmentes nach dem Stand der Technik;

Figur 2b: Struktur eines erfindungsgemäßen Audiosegmentes, das ein Phon mit nachgelagerten Koartikualtionsbereichen wiedergibt;

Figur 2c: Struktur eines erfindungsgemäßen Audiosegmentes, das ein Phon mit vorgelagerten Koartikualtionsbereichen wiedergibt;

Figur 2d: Struktur eines erfindungsgemäßen Audiosegmentes, das ein Phon mit nachgelagerten Koartikualtionsbereichen wiedergibt und eventuell vom Konkatenationsverfahren (z.B. Crossfade) benötigte zusätzliche Bereiche enthält;

Figur 2e: Struktur eines erfindungsgemäßen Audiosegmentes, das ein Phon mit vorgelagerten Koartikualtionsbereichen wiedergibt und eventuell vom Konkatenationsverfahren (z.B. Crossfade) benötigte zusätzliche Bereiche enthält;

Figur 2f: Strukturen von erfindungsgemäßen Audiosegmenten, das ein Polyphon mit jeweils nachgelagerten Koartikualtionsbereichen wiedergeben;

Figur 2g: Struktur eines erfindungsgemäßen Audiosegmentes, das ein Polyphon mit jeweils vorgelagerten Koartikualtionsbereichen wiedergibt;

Figur 2h: Struktur eines erfindungsgemäßen Audiosegmentes, das ein Polyphon mit jeweils nachgelagerten Koartikualtionsbereichen wiedergibt und eventuell vom Konkatenationsverfahren (z.B. Crossfade) benötigte zusätzliche Bereiche enthält;

Figur 2i: Struktur eines erfindungsgemäßen Audiosegmentes, das ein Polyphon mit jeweils vorgelagerten Koartikualtionsbereichen wiedergibt und eventuell vom Konkatenationsverfahren (z.B. Crossfade) benötigte zusätzliche Bereiche enthält;

Figur 2j: Strukturen von erfindungsgemäßen Audiosegmenten, das einen Teil eines Phons oder Phone vom Anfang einer Phonfolge wiedergeben;

Figur 2k: Struktur eines erfindungsgemäßen Audiosegmentes, das Phone vom Ende einer Phonfolge wiedergibt;

Figur 3a: Konkatenation gemäß dem Stand der Technik am Beispiel zweier Audiosegmente;

Figur 3b: Konkatenation gemäß dem erfindungsgemäßen Verfahren am Beispiel zweier Audiosegmente, die je ein Phon mit nachgelagerten Koartikulationsbereichen enthalten, mittels eines Crossfades (Fig. 3bI) und eines Hardfades (Fig. 3bII), wobei das erste Phon vom Anfang einer Lauteinheitenfolge stammt;

Figur 3c: Konkatenation gemäß dem erfindungsgemäßen Verfahren am Beispiel zweier Audiosegmente, die je ein Phon mit nachgelagerten Koartikulationsbereichen enthalten, mittels eines Crossfades (Fig. 3cI) und eines Hardfades (Fig. 3cII);

Figur 3d: Konkatenation gemäß dem erfindungsgemäßen Verfahren am Beispiel zweier Audiosegmente, die je ein Phon mit vorgelagerten Koartikulationsbereichen enthalten, mittels eines Crossfades (Fig. 3dI) und eines Hardfades (Fig. 3dII);

Figur 3e: Konkatenation gemäß dem erfindungsgemäßen Verfahren am Beispiel zweier Audiosegmente, die je ein Phon mit nachgelagerten Koartikulationsbereichen enthalten, mittels eines Crossfades (Fig. 3eI) und eines Hardfades (Fig. 3eII), wobei das erste Phon vom Ende einer Lauteinheitenfolge stammt; und

Figur 4: Schematische Darstellung der Schritte eines erfindungsgemäßen Verfahrens zur Erzeugung synthetisierter akustischer Daten.

Die im folgenden benutzten Bezugszeichen beziehen sich auf die Figur 1 und die im folgenden für die verschiedenen Verfahrensschritte benutzten Nummern beziehen sich auf die Figur 4.

Um mit Hilfe der Erfindung beispielsweise einen Text in synthetisierte Sprache umzuwandeln, ist es notwendig in einem vorgelagerten Schritt diesen Text in eine Folge von Phonemen unter Verwendung bekannter Verfahren oder Vorrichtungen zu unterteilen. Vorzugsweise sind auch dem Text entsprechende prosodische Informationen zu erzeugen. Die Phonemfolge sowie die prosodischen Informationen dienen als Eingabegrößen für das erfindungsgemäße Verfahren bzw. die erfindungsgemäße Vorrichtung.

Die zu synthetisierenden Phoneme werden einer Eingabeeinheit 101 der Vorrichtung 1 zur Erzeugung synthetisierter Sprachdaten zugeführt und in einer ersten Speichereinheit 103 abgelegt (siehe Figur 1). Mit Hilfe einer Auswahleinrichtung 103 werden aus einem Audiosegmente (Elemente) enthaltenden Inventar, das in einer Datenbank 107 gespeichert ist, die Audiosegmente ausgewählt, die Phone oder Teile von Phonen wiedergeben, die den einzelnen eingegebenen Phonemen oder Teilen davon entsprechen und in einer Reihenfolge, die der Reihenfolge der eingegebenen Phoneme entspricht, in einer zweiten Speichereinheit 104 gespeichert . Falls das Inventar Polyphone wiedergebende Audiosegmente enthält, so wählt die Auswahleinrichtung 103 vorzugsweise die Audiosegmente aus, die die längsten Polyphone wiedergeben, die einer Folge von Phonemen aus der eingegebenen Phonemfolge entsprechen.

Stellt die Datenbank 107 ein Inventar mit Audiosegmenten unterschiedlicher Arten zur Verfügung, so wählt die Auswahleinrichtung 103 vorzugsweise die längsten Audiosegmente aus, die den Phonemfolgen oder Teilen davon entsprechen, um die eingegebene Phonemfolge und/oder eine Folge von Phonemen aus einer minima-

len Anzahl von Audiosegmenten zu synthetisieren. Hierbei ist es vorteilhaft, verkettete Phone als Inventarelemente zu verwenden, die aus einem zeitlich vorgelagerten statischen Phon und einem zeitlich nachgelagerten dynamischen Phon bestehen. So entstehen Inventarelemente, die aufgrund der Einbettung der dynamischen Phone immer mit einem statischen Phon beginnen. Dadurch vereinfacht und vereinheitlicht sich das Vorgehen bei Konkatenationen solcher Inventarelemente, da hierfür nur Crossfades benötigt werden.

Um eine koartikulationsgerechte Konkatenation der zu verkettenden Audiosegmente zu erzielen, werden mit Hilfe einer Konkatenationseinrichtung 111 die Konkatenationsmomente zweier aufeinanderfolgender Audiosegmente wie folgt festgelegt:

- Soll ein Audiosegment zu Synthetisierung des Anfanges der eingegebenen Phonemfolge (Schritt 1) verwendet werden, so ist aus dem Inventar ein Audiosegment zu wählen, das einen Wortanfang wiedergibt und mit einem zeitlich nachgelagerten Audiosegment zu verketten (siehe Figur 3b und Schritt 3 in Figur 4).

- Bei der Konkatenation eines zweiten Audiosegmentes an ein zeitlich vorgelagertes erstes Audiosegment ist zu unterscheiden, ob das zweite Audiosegment mit einem statischen Phon oder einem dynamischen Phon beginnt, um die Wahl des Momentes der Konkatenation entsprechend zu treffen (Schritt 6).

- Beginnt das zweite Audiosegment mit einem statischen Phon, wird die Konkatenation in Form eines Crossfades durchgeführt, wobei der Moment der Konkatenation im zeitlich hinteren Bereich des ersten Audiosegmentes und im zeitlich vorderen Bereich des zweiten Audiosegmentes gelegt wird, wodurch sich diese beiden Bereiche bei der Konkatenation überlappen oder wenigstens unmittelbar aneinandergrenzen (siehe Figuren 3c und 3d, Konkatenation mittels Crossfade).

- Beginnt das zweite Audiosegment mit einem dynamischen Phon, wird die Konkatenation in Form eines Hardfades durchgeführt,

wobei der Moment der Konkatenation zeitlich unmittelbar hinter der zeitlich hinteren Bereich des ersten Audiosegmentes und zeitlich unmittelbar vor dem zeitlich vorderen Bereich des zweiten Audiosegmentes gelegt wird (siehe Figuren 3c und 3d, Konkatenation mittels Hardfade).

Auf diese Weise können aus diesen ursprünglich verfügbaren Audiosegmenten, die Phone oder Polyphone wiedergeben, neue Polyphone wiedergebende Audiosegmente erzeugt werden, die mit einem statischen Phon beginnen. Dies erreicht man, indem Audiosegmente, die mit einem dynamischen Phon beginnen, zeitlich nachgelagert mit Audiosegmenten, die mit einem statischen Phon beginnen, verkettet werden. Dies vergrößert zwar die Zahl der Audiosegmente bzw. den Umfang des Inventars, kann aber bei der Erzeugung synthetisierter Sprachdaten einen rechentechnischen Vorteil darstellen, da weniger einzelne Konkatenationen zur Erzeugung einer Phonemfolge erforderliche sind und Konkatenationen nur noch in Form eines Crossfades durchgeführt werden müssen. Vorzugsweise werden die so erzeugten neuen verketteten Audiosegmente der Datenbank 107 oder einer anderen Speichereinheit 113 zugeführt.

Ein weiterer Vorteil dieser Verkettung der ursprüngliche Audiosegmente zu neuen längeren Audiosegmenten ergibt sich, wenn sich beispielsweise eine Folge von Phonemen in der eingegebenen Phonemfolge häufig wiederholt. Dann kann auf eines der neuen entsprechend verketteten Audiosegmente zurückgegriffen werden und es ist nicht notwendig, bei jedem Auftreten dieser Folge von Phonemen eine erneute Konkatenation der ursprünglich vorhandenen Audiosegmente durchzuführen. Vorzugsweise sind bei der Speicherung solcher verketteten Audiosegmente auch übergreifende Koartikulationseffekte zu erfassen bzw. spezifische Koartikulationseffekte in Form zusätzlicher Daten dem gespeicherten verketteten Audiosegment zuzuordnen.

Soll ein Audiosegment zu Synthetisierung des Endes der eingegebenen Phonemfolge verwendet werden, so ist aus dem Inventar ein Audiosegment zu wählen, das ein Wortende wiedergibt und mit

einem zeitlich vorgelagertes Audiossegment zu verketten (siehe Figur 3e und Schritt 8 in Figur 4).

Die einzelnen Audiosegmente werden in der Datenbank 107 kodiert gespeichert, wobei die kodierte Form der Audiosegmente neben der Wellenform des jeweiligen Audiosegmentes angibt, welche(s) Phon(e) das jeweilige Audiosegment wiedergibt, welche Art der Konkatenation (z.B. Hardfade, linearer oder exponentieller Crossfade) mit welchem zeitlich nachfolgenden Audiosegment durchzuführen ist und zu welchem Moment die Konkatenation mit welchem zeitlich nachfolgenden Audiosegment stattfindet. Vorzugsweise enthält die kodierte Form der Audiosegmente auch Informationen bezüglich der Prosodie und übergeordneten Koartikulationen, die bei einer Synthetisierung der gesamten vom Sprecher aufgenommene Phonemfolge und/oder Folgen von Phonem verwendet werden, um eine zusätzliche Verbesserung der Sprachqualität zu erzielen.

Bei der Wahl der Audiosegmente zur Synthetisierung der eingegebenen Phonemfolge werden als zeitlich nachgelagerte Audiosegmente solche gewählt, die den Eigenschaften der jeweils zeitlich vorgelagerten Audiosegmente, d.h. Konkatenationsart und Konkatenationsmoment, entsprechen. Nachdem die der Phonemfolge entsprechenden Audiosegmente aus der Datenbank 107 gewählt wurden, erfolgt die Verkettung zweier aufeinanderfolgender Audiosegmente mit Hilfe der Konkatenationseinrichtung 111 folgendermaßen. Es wird die Wellenform, die Konkatenationsart und der Konkatenationsmoment des ersten Audiosegmentes und des zweiten Audiosegmentes aus der Datenbank (Figur 3a und Schritt 10 und 11) geladen. Vorzugsweise werden bei der oben erwähnten Wahl der Audiosegmente solche Audiosegmente gewählt, die hinsichtlich ihrer Konkatenationsart und ihres Konkatenationsmoment zu einander passen. In diesem Fall ist das Laden der Informationen bezüglich der Konkatenationsart und des Konkatenationsmomentes des zweiten Audiosegmentes ist nicht mehr notwendig.

Zur Konkatenation der beiden Audiosegmente werden die Wellenform des ersten Audiosegmentes in einem zeitlich hinteren Bereich und die Wellenform des zweiten Audiosegmentes in einem zeitlich vorderen Bereich jeweils mit einer geeigneten Gewichtungsfunktion multipliziert (siehe Figur 3a, Schritt 12 und 13). Die Längen des zeitlich hinteren Bereiches des ersten Audiosegmentes und des zeitlich vorderen Bereiches des zweiten Audiosegmentes ergeben sich aus der Konkatenationsart und zeitlichen Lage des Konkatenationsmomentes, wobei diese Längen auch in der kodierten Form der Audiosegmente in der Datenbank gespeichert werden können.

Sind die beiden Audiosegmente mit einem Crossfade zu verketten, werden diese entsprechend dem jeweiligen Konkatenationsmoment überlappend addiert (siehe Figuren 3c und 3d, Schritt 15). Vorzugsweise ist hierbei ein linearer symmetrischer Crossfade zu verwenden, es kann aber auch jede andere Art eines Crossfades eingesetzt werden. Ist eine Konkatenation in Form eines Hardfades durchzuführen, werden die beiden Audiosegmente nicht überlappend hintereinander verbunden (siehe Figur 3c und 3d, Schritt 15). Wie in Figur 3d zu sehen ist, werden hierbei die beiden Audiosegmente zeitlich unmittelbar hintereinander angeordnet. Um die so erzeugten synthetisierten Sprachdaten weiterverarbeiten zu können, werden diese vorzugsweise in einer dritten Speichereinheit 115 abgelegt.

Für die weitere Verkettung mit nachfolgenden Audiosegmenten werden die bisher verketteten Audiosegmente als erstes Audiosegment betrachtet (Schritt 16) und der oben beschriebenen Verkettungsprozeß solange wiederholt, bis die gesamte Phonemfolge synthetisiert wurde.

Zur Verbesserung der Qualität der synthetisierten Sprachdaten sind vorzugsweise auch die prosodischen Informationen, die zusätzlich zu der Phonemfolge eingegeben werden, bei der Verkettung der Audiosegmente zu berücksichtigen. Mit Hilfe bekannter Verfahren kann die Frequenz, Dauer und Amplitude der Audiosegmente vor und/oder nach deren Konkatenation so verän-

dert werden, daß die synthetisierten Sprachdaten eine natürliche Wort- und/oder Satzmelodie aufweisen (Schritte 14, 17 oder 18). Hierbei ist es zu bevorzugen, Konkatenationsmomente an Nullstellen der Audiosegmente zu wählen.

Um die Übergänge zwischen zwei aufeinander folgenden Audiosegmenten zu optimieren, ist zusätzlich die Anpassung der Frequenzen, Dauer und Gesamtamplituden sowie von Amplituden in verschiedenen Frequenzbereichen der beiden Audiosegmente im Bereich des Konkatenationsmomentes vorgesehen. Des weiteren erlaubt es die Erfindung, auch übergeordnete akustische Phänomene einer realen Sprache, wie z.B. übergeordnete Koartikulationseffekte oder Sprachstil (u.a. Flüstern, Betonung, Gesangsstimme oder Falsett) bei der Synthetisierung der Phonemfolgen zu berücksichtigen. Hierfür werden Informationen, die solche übergeordnete Phänomene betreffen, zusätzlich in kodierter Form mit den entsprechenden Audiosegmenten gespeichert, um so bei der Auswahl der Audiosegmente nur solche zu wählen, die den übergeordneten Koartikulationseigenschaften der zeitlich vor- und/oder nachgelagerten Audiosegmente entsprechen.

Die so erzeugten synthetisierten Sprachdaten haben vorzugsweise eine Form, die es unter Verwendung einer Ausgabeeinheit 117 erlaubt, die Sprachdaten in akustische Sprachsignale umzuwandeln und die Sprachdaten und/oder Sprachsignale auf einem akustischen, optischen oder elektrischen Datenträger zu speichern (Schritt 19).

Mit dieser Erfindung ist es erstmals möglich synthetisierte Sprachsignale durch eine koartikulationsgerechte Konkatenation einzelner Audiosegmente zu erzeugen, da der Moment der Konkatenation in Abhängigkeit der jeweils zu verkettenden Audiosegmente gewählt wird. Auf diese Weise kann eine synthetisierte Sprache erzeugt werden, die vom einer natürlichen Sprache nicht mehr zu unterscheiden ist. Im Gegensatz zu bekannten Verfahren oder Vorrichtungen werden die hier verwendeten Audiosegmente nicht durch ein Einsprechen ganzer Worte erzeugt, um eine authentische Sprachqualität zu gewährleisten. Daher ist es mit

dieser Erfindung möglich, synthetisierte Sprache beliebigen Inhalts in der Qualität einer real gesprochenen Sprache zu erzeugen.

Obwohl diese Erfindung am Beispiel der Sprachsynthese beschrieben wurde, ist die Erfindung nicht auf den Bereich der synthetisierten Sprache beschränkt, sondern kann zu Synthetisierung beliebiger akustischer Daten verwendet werden. Daher ist diese Erfindung auch für eine Erzeugung und/oder Bereitstellung von synthetisierten Sprachdaten und/oder Sprachsignale für beliebige Sprachen oder Dialekte sowie zur Synthese von Musik einsetzbar.

Patentansprüche

1. Verfahren zur Erzeugung synthetisierter akustischer Daten, die aus einer Folge von Lauteinheiten bestehen, durch Konkatenation von Audiosegmenten, mit folgenden Schritten:
- Auswahl von wenigstens zwei Audiosegmenten, die Lauteinheiten wiedergeben, aus einer Datenbank zu synthetisierender akustischer Daten, dadurch gekennzeichnet, daß
- jedes Audiosegment wenigstens einen Solo-Artikulationsbereich aufweist, und
- der Moment der Konkatenation eines Anfangs eines verwendeten Teiles eines zweiten Audiosegmentes mit dem Ende eines verwendeten Teiles eines ersten Audiosegment in Abhängigkeit von Eigenschaften des verwendeten Teiles des zweiten Audiosegmentes in einen Bereich gelegt wird, der zeitlich unmittelbar vor dem verwendeten Teil des zweiten Audiosegmentes beginnt und nach dem zeitlich ersten verwendeten Solo-Artikulationsbereich des verwendeten Teiles des zweiten Audiosegmentes endet.

2. Verfahren nach Anspruch 1, dadurch gekennzeichnet, daß für die Erzeugung der synthetisierten akustischen Daten, die einer ersten Lauteinheit am Anfang der Lauteinheitenfolge entsprechen, ein solches Audiosegment ausgewählt wird, dessen zeitlich vorderer Bereich des verwendeten Teiles des Audiosegmentes die Eigenschaften des Anfangs der Lauteinheitenfolge aufweist.

3. Verfahren nach einem der Ansprüche 1 oder 2, dadurch gekennzeichnet, daß der Moment der Konkatenation des zweiten Audiosegmentes mit dem ersten Audiosegment so gewählt wird, daß er in der Umgebung der Grenzen des ersten verwendeten Solo-Artikulationsbereiches des verwendeten Teiles des zweiten Audiosegmentes liegt, wenn der verwendete Teil des zweiten Audiosegmentes mit einer statischen Lauteinheit beginnt, wobei ein zeitlich hinterer Bereich des verwendeten Teiles des ersten Audiosegmentes und ein zeitlich vorderer Bereich des verwendeten Teiles des zweiten Audiosegmentes gewichtet und danach beide Bereiche addiert werden (Crossfade), wobei die Länge

eines Überlappungsbereiches der beiden Bereiche in Abhängigkeit der zu synthetisierenden akustischen Daten bestimmt wird.

4. Verfahren nach einem der Ansprüche 1 bis 3, dadurch gekennzeichnet, daß der Moment der Konkatenation des zweiten Audiosegmentes mit dem ersten Audiosegment so gewählt wird, daß er zeitlich unmittelbar vor dem verwendeten Teil des zweiten Audiosegmentes liegt, wenn der verwendete Teil des zweiten Audiosegmentes mit einer dynamischen Lauteinheit beginnt, wobei ein zeitlich hinterer Bereich des verwendeten Teiles des ersten Audiosegmentes und ein zeitlich vorderer Bereich des verwendeten Teiles des zweiten Audiosegmentes gewichtet werden (Hardfade).

5. Verfahren nach einem der Ansprüche 1 bis 4, dadurch gekennzeichnet, daß für die Erzeugung der synthetisierten akustischen Daten, die einer letzten Lauteinheit am Ende der Lauteinheitenfolge entsprechen, ein solches Audiosegment ausgewählt wird, dessen zeitlich hinterer Bereich des verwendeten Teiles des Audiosegmentes die Eigenschaften des Endes der Lauteinheitenfolge aufweist.

6. Verfahren nach Anspruch 4, dadurch gekennzeichnet, daß weitere Audiosegmente dadurch gebildet werden, indem Audiosegmente, deren verwendeter Teil mit einer dynamischen Lauteinheit beginnt, oder eine Folge von Audiosegmenten, deren verwendete Teile mit dynamischen Lauteinheiten beginnen, mit wenigstens einem zeitlich vorgelagerten Audiosegment, dessen verwendeter Teil mit einer statischen Lauteinheit beginnt, verkettet werden.

7. Verfahren nach einem der Ansprüche 1 bis 6, dadurch gekennzeichnet, daß die zu synthetisierenden Sprachdaten in Gruppen von aufeinanderfolgenden Lauteinheiten zusammengefaßt werden, die jeweils durch ein einzelnes Audiosegment beschrieben werden.

8. Verfahren nach einem der Ansprüche 1 bis 7, dadurch gekennzeichnet, daß bei der Konkatenation eines zweiten Audiosegmentes mit einem ersten Audiosegment aus der Datenbank für das zweite Audiosegment ein Audiosegment gewählt wird, das die meisten aufeinanderfolgenden Lauteinheiten der zu synthetisierenden Daten wiedergibt, um bei der Erzeugung der synthetisierten Daten die minimale Anzahl von Audiosegmenten zu verwenden.

9. Verfahren nach einem der Ansprüche 1 bis 8, dadurch gekennzeichnet, daß eine Variation der Frequenz, der Dauer und der Gesamtamplitude der verwendeten Teile einzelner Audiosegmente sowie deren Amplitude in verschiedenen Frequenzbereichen in Abhängigkeit von Eigenschaften der Lauteinheitenfolge durchgeführt wird.

10. Verfahren nach einem der Ansprüche 1 bis 9, dadurch gekennzeichnet, daß eine Variation der Frequenz, der Dauer und der Gesamtamplitude der verwendeten Teile einzelner Audiosegmente sowie deren Amplitude in verschiedenen Frequenzbereichen in einem Bereich durchgeführt wird, in dem der Moment der Konkatenation liegt.

11. Verfahren nach einem der Ansprüche 1 bis 10, dadurch gekennzeichnet, daß der Moment der Konkatenation bei einer Nullstelle in den verwendeten Teilen des ersten und/oder des zweiten Audiosegmentes gewählt wird.

12. Verfahren nach einem der Ansprüche 1 bis 11, dadurch gekennzeichnet, daß die Auswahl der verwendeten Teile einzelner Audiosegmente, deren Variation sowie deren Konkatenation zusätzlich unter Verwendung heuristischen Wissens durchgeführt wird, das durch ein zusätzlich durchgeführtes heuristisches Verfahren gewonnen wird.

13. Verfahren nach einem der Ansprüche 1 bis 12, dadurch gekennzeichnet, daß eine Umwandlung der synthetisierten akustischen Daten in akustische Signale durchgeführt wird.

14. Verfahren nach einem der Ansprüche 1 bis 13, dadurch gekennzeichnet, daß die synthetisierten akustischen Daten auf einem Datenträger gespeichert werden.

15. Verfahren einem der Ansprüche 1 bis 14, dadurch gekennzeichnet, daß
- die zu synthetisierenden akustischen Daten Sprachdaten und die Lauteinheiten Phone sind,
- die statischen Lauteinheiten Vokale, Diphtonge, Liquide, Vibranten, Frikative und Nasale umfassen, und
- die dynamischen Lauteinheiten Plosive, Affrikate, Glottalstops und geschlagenen Laute umfassen.

16. Vorrichtung zur Erzeugung synthetisierter akustischer Daten, die aus einer Folge von Lauteinheiten bestehen, durch Konkatenation von Audiosegmenten, mit:
- einer Datenbank, in der die Audiosegmente der zu synthetisierender Daten gespeichert sind,
- einer Einrichtung zur Auswahl von wenigstens zwei die Lauteinheiten wiedergebenden Audiosegmenten aus der Datenbank, und
- einer Einrichtung zur Konkatenation der Audiosegmente, dadurch gekennzeichnet, daß
- die Datenbank Audiosegmente enthält, die wenigstens einen Solo-Artikulationsbereich aufweisen, und
- die Konkatenationeinrichtung geeignet ist, den Moment der Konkatenation eines Anfangs eines verwendeten Teils eines zweiten Audiosegmentes mit dem Ende eines verwendeten Teils eines ersten Audiosegmentes in Abhängigkeit von Eigenschaften des verwendeten Teils des zweiten Audiosegmentes in einen Bereich zu legen, der zeitlich unmittelbar vor dem verwendeten Teil des zweiten Audiosegmentes beginnt und nach dem zeitlich ersten verwendeten Solo-Artikulationsbereich des verwendeten Teils des zweiten Audiosegmentes endet.

17. Vorrichtung nach Anspruch 16, dadurch gekennzeichnet, daß die Datenbank Audiosegmente enthält, deren verwendete Teile am Anfang einer Lauteinheitenfolge auftretende Lauteinheiten wiedergeben.

18. Vorrichtung nach einem der Ansprüche 16 oder 17, dadurch gekennzeichnet, daß die Konkatenationeinrichtung zusätzlich umfaßt:
- Einrichtungen zur Konkatenation eines ersten Audiosegmentes mit einem zweiten Audiosegment, dessen verwendeter Teil mit einer statischen Lauteinheit beginnt, im Bereich der Grenzen des ersten verwendeten Solo-Artikulationsbereiches des verwendeten Teils des zweiten Audiosegmentes,
- Einrichtungen zur Gewichtung eines zeitlich hinteren Bereiches des verwendeten Teils des ersten Audiosegmentes und eines zeitlich vorderen Bereiches des verwendeten Teils des zweiten Audiosegmentes, und
- Einrichtungen zur Addition der beiden Bereiche.

19. Vorrichtung nach einem der Ansprüche 16 bis 18, dadurch gekennzeichnet, daß die Konkatenationeinrichtung zusätzlich umfaßt:
- Einrichtungen zur Konkatenation eines ersten Audiosegmentes mit einem zweiten Audiosegment, dessen verwendeter Teil mit einer dynamischen Lauteinheit beginnt, zeitlich unmittelbar vor dem verwendeten Teil des zweiten Audiosegmentes, und
- Einrichtungen zur Gewichtung eines zeitlich hinteren Bereiches des verwendeten Teil des ersten Audiosegmentes und eines zeitlich vorderen Bereiches des verwendeten Teil des zweiten Audiosegmentes.

20. Vorrichtung nach einem der Ansprüche 16 bis 19, dadurch gekennzeichnet, daß die Datenbank Audiosegmente enthält, deren verwendete Teile am Ende einer Lauteinheitenfolge auftretende Lauteinheiten wiedergeben.

21. Vorrichtung nach einem der Ansprüche 16 bis 22, dadurch gekennzeichnet, daß die Datenbank eine Gruppe von Audiosegmenten enthält, deren verwendete Teile mit einer statischen Lauteinheit beginnen.

22. Vorrichtung nach einem der Ansprüche 16 bis 21, dadurch gekennzeichnet, daß die Konkatenationseinrichtung zusätzlich umfaßt:
- eine Einrichtung zur Erzeugung weiterer Audiosegmente durch Konkatenation von Audiosegmenten, deren verwendete Teile mit einer statischen Lauteinheit beginnen, mit zeitlich nachgelagerten Audiosegmenten, deren verwendete Teile mit einer dynamischen Lauteinheit beginnen, und
- eine Einrichtung, die die weiteren Audiosegmente der Datenbank oder der Auswahleinrichtung zuführt.

23. Vorrichtung nach einem der Ansprüche 16 bis 22, dadurch gekennzeichnet, daß die Datenbank eine Gruppe von Audiosegmenten enthält, die jeweils eine Folge von Lauteinheiten wiedergeben.

24. Vorrichtung nach einem der Ansprüche 16 bis 23, dadurch gekennzeichnet, daß die Auswahleinrichtung geeignet ist, bei der Auswahl der Audiosegmente aus der Datenbank, die Audiosegmente auszuwählen, die die meisten aufeinanderfolgenden Lauteinheiten der zu synthetisierenden Daten wiedergeben, um bei der Erzeugung der synthetisierten Daten die minimal Anzahl von Audiosegmenten zu verwenden.

25. Vorrichtung nach einem der Ansprüche 16 bis 24, dadurch gekennzeichnet, daß die Konkatenationseinrichtung zusätzlich eine Einrichtung zur Variation der Frequenz, der Dauer und der Gesamtamplitude der verwendeten Teile einzelner Audiosegmente sowie deren Amplitude in verschiedenen Frequenzbereichen in Abhängigkeit von Eigenschaften der Lauteinheitenfolge umfaßt.

26. Vorrichtung nach einem der Ansprüche 16 bis 25, dadurch gekennzeichnet, daß die Konkatenationseinrichtung zusätzlich eine Einrichtung zur Variation der Frequenz, der Dauer und der Gesamtamplitude der verwendeten Teile einzelner Audiosegmente sowie deren Amplitude in verschiedenen Frequenzbereichen in einem Bereich durchgeführt wird, in dem der Moment der Konkatenation liegt, umfaßt.

27. Vorrichtung nach einem der Ansprüche 16 bis 26, dadurch gekennzeichnet, daß die Konkatenationseinrichtung zusätzlich eine Einrichtung zur Auswahl des Momentes der Konkatenation bei einer Nullstelle in den verwendeten Teilen des ersten und/oder des zweiten Audiosegmentes aufweist.

28. Vorrichtung nach einem der Ansprüche 16 bis 27, dadurch gekennzeichnet, daß die Auswahleinrichtung zusätzlich eine Einrichtung zur Implementation heuristischen Wissens umfaßt, das die Auswahl der einzelnen Audiosegmente, deren Variation sowie die Konkatenation der Audiosegmente betrifft.

29. Vorrichtung nach einem der Ansprüche 16 bis 28, dadurch gekennzeichnet, daß zusätzlich Einrichtungen zur Umwandlung der synthetisierten akustischen Daten in akustische Signale vorgesehen sind.

30. Vorrichtung nach einem der Ansprüche 16 bis 29, dadurch gekennzeichnet, daß zusätzlich Einrichtungen zur Speicherung der synthetisierten akustischen Daten auf einem Datenträger vorgesehen sind.

31. Vorrichtung nach einem der Ansprüche 16 bis 30, dadurch gekennzeichnet, daß
- die Datenbank Audiosegmente enthält, die jeweils wenigstens einen Teil eines Phons wiedergeben, wobei eine statische Lauteinheit Vokale, Diphtonge, Liquide, Vibranten, Frikative und Nasale umfaßt und
eine dynamische Lauteinheit Plosive, Affrikate, Glottalstops und geschlagene Laute umfaßt, und
- die Konkatenationseinrichtung geeignet ist, die Audiosegmente zu synthetisierten Sprachdaten zu verketten.

32. Synthetisierte Sprachsignale, die aus einer Folge von Phonen bestehen, wobei die Sprachsignale erzeugt werden, indem:
- wenigstens zwei die Phone wiedergebende Audiosegmente aus einer Datenbank ausgewählt werden, und

- die Audiosegmente durch eine Konkatenation verkettet werden, wobei
- jedes Audiosegment wenigstens einen Solo-Artikulationsbereich aufweist, und
- der Moment der Konkatenation des Anfangs eines verwendeten Teiles eines zweiten Audiosegementes mit dem Ende eines verwendeten Teiles eines ersten Audiosegmentes in Abhängigkeit von Eigenschaften des verwendeten Teiles des zweiten Audiosegmentes in einen Bereich gelegt wird, der zeitlich unmittelbar vor dem verwendeten Teil des zweiten Audiosegmentes beginnt und nach dem zeitlich ersten verwendeten Solo-Artikulationsbereich des verwendeten Teiles des zweiten Audiosegementes endet.

33. Synthetisierte Sprachsignale nach Anspruch 32, dadurch gekennzeichnet, daß das erste Phon in der Phonfolge durch ein Audiosegment erzeugt wird, dessen verwendeter Teil einen zeitlich vorderen Bereich hat, der die Eigenschaften des Anfangs der Phonfolge aufweist.

34. Synthetisierte Sprachsignale nach einem der Ansprüche 32 oder 33, dadurch gekennzeichnet, daß die Sprachsignale erzeugt werden, indem
- das erste Audiosegment und das zweite Audiosegment zu einem Moment konkateniert werden, der in der Umgebung der Grenzen des ersten verwendeten Solo-Artikulationsbereiches des verwendeten Teiles des zweiten Audiosegmentes liegt, wenn der verwendete Teil des zweite Audiosegment mit einem statischen Phon beginnt, wobei ein statischer Phon ein Vokal, ein Diphtong, ein Liquid, ein Frikativ, ein Vibrant oder ein Nasal sein kann, und
- ein zeitlich hinterer Bereich des verwendeten Teiles des ersten Audiosegmentes und ein zeitlich vorderer Bereich des verwendeten Teiles des zweiten Audiosegmentes gewichtet und beide Bereiche addiert werden (Crossfade).

35. Synthetisierte Sprachsignale nach einem der Ansprüche 32 bis 34, dadurch gekennzeichnet, daß die Sprachsignale erzeugt werden, indem
- das erste Audiosegment und das zweite Audiosegment zu einem Moment konkateniert werden, der zeitlich unmittelbar vor dem verwendeten Teil des zweiten Audiosegmentes liegt, wenn der verwendete Teil des zweiten Audiosegmentes mit einem dynamischen Phon beginnt, wobei ein dynamischer Phon
ein Plosiv, ein Affrikat, ein Glottalstop oder ein geschlagener Laut sein kann, und
- ein zeitlich hinterer Bereich des verwendeten Teiles des ersten Audiosegmentes und ein zeitlich vorderer Bereich des verwendeten Teiles des zweiten Audiosegmentes gewichtet werden (Hardfade).

36. Synthetisierte Sprachsignale nach einem der Ansprüche 32 bis 35, dadurch gekennzeichnet, daß das letzte Phon in der Phonfolge durch ein Audiosegment erzeugt wird, dessen verwendeter Teil einen zeitlich hinteren Bereich hat, der die Eigenschaften des Endes der Phonfolge aufweist.

37. Synthetisierte Sprachsignale nach Anspruch 36, dadurch gekennzeichnet, daß die Sprachsignale durch eine Konkatenation eines ersten Audiosegmentes mit einem zweiten Audiosegment erzeugt werden, wobei der verwendete Teil des zweiten Audiosegmentes einen mit einem statischen Phon beginnenden und zeitlich vorgelagerten Bereich und wenigstens einen dem verwendeten Teil zeitlich nachgelagerten Bereich umfaßt, der mit einem dynamischen Phon beginnt.

38. Synthetisierte Sprachsignale nach einem der Ansprüche 32 bis 36, dadurch gekennzeichnet, daß die Sprachsignale durch Konkatenation von Audiosegmenten erzeugt werden, die Polyphone wiedergeben.

39. Synthetisierte Sprachsignale nach einem der Ansprüche 32 bis 36, dadurch gekennzeichnet, daß zur Erzeugung der Sprachsignale aus der Datenbank die Audiosegmente ausgewählt werden, die die meisten zusammenhängenden Phone der Folge der Phone wiedergeben, um bei der Erzeugung der Sprachsignale die minimal Anzahl von Audiosegmenten zu verwenden.

40. Synthetisierte Sprachsignale nach einem der Ansprüche 32 bis 39, dadurch gekennzeichnet, daß die Sprachsignale durch Konkatenation der verwendeten Teile von Audiosegmenten erzeugt werden, deren Frequenz, Dauer und Gesamtamplitude sowie deren Amplituden in verschiedenen Frequenzbereichen in Abhängigkeit von Eigenschaften der Phonfolge variiert werden.

41. Synthetisierte Sprachsignale einem der Ansprüche 32 bis 40, dadurch gekennzeichnet, daß die Sprachsignale durch Konkatenation von Audiosegmenten erzeugt werden, deren Frequenz, Dauer, Gesamtamplitude und deren Amplituden in verschiedenen Frequenzbereichen der jeweils verwendeten Teile der Audiosegmente in einem Bereich variiert werden, in dem der Moment der Konkatenation liegt.

42. Synthetisierte Sprachsignale einem der Ansprüche 32 bis 41, dadurch gekennzeichnet, daß der Moment der Konkatenation bei einer Nullstelle in den verwendeten Teilen des ersten und/oder des zweiten Audiosegmentes liegt.

43. Synthetisierte Sprachsignale nach einem der Ansprüche 32 bis 42, dadurch gekennzeichnet, daß die Sprachsignale geeignet sind, in akustische Signale umgewandelt zu werden.

44. Synthetisierte Sprachsignale nach den Ansprüchen 32 bis 43, dadurch gekennzeichnet, daß die Sprachsignale geeignet sind, auf einem Datenträger gespeichert zu werden.

45. Datenträger, der ein Computerprogramm zur Erzeugung von synthetisierten akustischen Daten, die aus einer Folge von Lauteinheiten bestehen, durch Konkatenation von Audiosegmenten enthält, wobei das Computerprogramm folgende Schritte ausführt:
- Auswahl von wenigstens zwei die Lauteinheiten wiedergebenden Audiosegmenten aus einer Datenbank zu synthetisierender akustischer Daten, und
- Konkatenation der Audiosegmente, wobei
- jedes Audiosegment wenigstens einen Solo-Koartikulationsbereich aufweist, und
- der Moment der Konkatenation des Anfangs eines verwendeten Teiles eines zweiten Audiosegmentes mit dem Ende eines verwendeten Teiles eines ersten Audiosegmentes in Abhängigkeit von Eigenschaften des verwendeten Teiles des zweiten Audiosegmentes in einen Bereich gelegt wird, der zeitlich unmittelbar vor dem verwendeten Teil des zweiten Audiosegmentes beginnt und nach dem zeitlich ersten verwendeten Solo-Koartikulationbereich des verwendeten Teiles des zweiten Audiosegmentes endet.

46. Datenträger nach Anspruch 45, dadurch gekennzeichnet, daß das Computerprogramm zur Erzeugung der synthetisierten akustischen Daten, die einer ersten Lauteinheit am Anfang der Lauteinheitenfolge entsprechen, ein solches Audiosegment auswählt, dessen verwendeter Teil einen zeitlich vorderen Bereich hat, der die Eigenschaften des Anfangs der Lauteinheitenfolge aufweist.

47. Datenträger nach einem der Ansprüche 45 oder 46, dadurch gekennzeichnet, daß das Computerprogramm den Moment der Konkatenation des zweiten Audiosegmentes mit dem ersten Audiosegment so wählt, daß er in der Umgebung der Grenzen des ersten verwendeten Solo-Artikulationsbereiches des verwendeten Teiles des zweiten Audiosegmentes liegt, wenn der verwendete Teil des zweiten Audiosegmentes mit einer statischen Lauteinheit beginnt, und einen zeitlich hinteren Bereich des verwendeten Teiles des ersten Audiosegmentes und einen zeitlich vorderen Bereich des verwendeten Teiles des zweiten Audiosegmentes wichtet und beide Bereiche addiert.

48. Datenträger nach einem der Ansprüche 45 bis 47, dadurch gekennzeichnet, daß das Computerprogramm den Moment der Konkatenation des zweiten Audiosegmentes mit dem ersten Audiosegment so wählt, daß er zeitlich unmittelbar vor dem verwendeten Teil des zweiten Audiosegmentes liegt, wenn der verwendete Teil des zweiten Audiosegmentes mit einer dynamischen Lauteinheit beginnt, und einen zeitlich hinteren Bereich des verwendeten Teiles des ersten Audiosegmentes und einen zeitlich vorderen Bereich des verwendeten Teiles des zweiten Audiosegmentes wichtet.

49. Datenträger nach einem der Ansprüche 45 bis 48, dadurch gekennzeichnet, daß das Computerprogramm zur Erzeugung der synthetisierten akustischen Daten, die einer letzten Lauteinheit am Ende der Lauteinheitenfolge entsprechen, ein solches Audiosegment auswählt, dessen verwendeter Teil einen zeitlich hinteren Bereich hat, der die Eigenschaften des Endes der Lauteinheitenfolge aufweist.

50. Datenträger nach einem der Ansprüche 45 bis 49, dadurch gekennzeichnet, daß das Computerprogramm bei der Konkatenation eines zweiten Audiosegmentes mit einem ersten Audiosegment aus der Datenbank für das zweite Audiosegment ein Audiosegment wählt, das die meisten aufeinanderfolgenden Lauteinheiten der zu synthetisierenden Daten wiedergibt, um bei der Erzeugung der synthetisierten Daten die minimal Anzahl von Audiosegmenten zu verwenden.

51. Datenträger nach einem der Ansprüche 45 bis 50, dadurch gekennzeichnet, daß das Computerprogramm eine Variation der Frequenz, Dauer und Gesamtamplitude der verwendeten Teile einzelner Audiosegmente und deren Amplituden in verschiedenen Frequenzbereichen in Abhängigkeit von Eigenschaften der Lauteinheitenfolge durchführt.

52. Datenträger nach einem der Ansprüche 45 bis 51, dadurch gekennzeichnet, daß das Computerprogramm eine Variation der Frequenz, Dauer und Gesamtamplitude der verwendeten Teile einzelner Audiosegmente und deren Amplituden in verschiedenen Frequenzbereichen in einem Bereich durchführt, in dem der Moment der Konkatenation liegt.

53. Datenträger nach einem der Ansprüche 45 bis 52, dadurch gekennzeichnet, daß Computerprogramm den Moment der Konkatenation bei einer Nullstelle in den verwendeten Teilen des ersten und/oder des zweiten Audiosegmentes festlegt.

54. Datenträger nach einem der Ansprüche 45 bis 53, dadurch gekennzeichnet, daß das Computerprogramm eine Implementation von heuristischem Wissen durchführt, das die Auswahl der einzelnen Audiosegmente, deren Variation sowie die Konkatenation der Audiosegmente betrifft.

55. Datenträger nach einem der Ansprüche 45 bis 54, dadurch gekennzeichnet, daß das Computerprogramm die synthetisierten akustischen Daten in akustische umwandelbare Daten umwandelt.

56. Datenträger nach einem der Ansprüche 45 bis 55, dadurch gekennzeichnet, daß das Computerprogramm die synthetisierten akustischen Daten auf einem Datenträger speichert.

57. Datenträger nach einem der Ansprüche 45 bis 56, dadurch gekennzeichnet, daß das Computerprogramm zur Erzeugung synthetisierter Sprachdaten geeignet ist, wobei die Lauteinheiten Phone sind, die statischen Lauteinheiten Vokale, Diphtonge, Liquide, Vibranten, Frikative und Nasale und die dynamischen Lauteinheiten Plosive, Affrikate, Glottalstops und geschlagene Laute umfassen.

58. Akustischer, optischer oder elektrischer Datenspeicher, der Audiosegmente enthält, die jeweils wenigstens einen Solo-Artikualtionsbereich aufweisen, um durch eine Konkatenation von verwendeten Teile der Audiosegmente unter Verwendung des Ver-

fahrens nach Anspruch 1 oder der Vorrichtung nach Anspruch 16 oder des Datenträgers nach Anspruch 45 synthetisierte akustische Daten zu erzeugen.

59. Datenspeicher nach Anspruch 58, dadurch gekennzeichnet, daß eine Gruppe der Audiosegmente Phone oder Teile von Phonen wiedergeben.

60. Datenspeicher nach einem der Ansprüche 58 oder 59, dadurch gekennzeichnet, daß eine Gruppe der Audiosegmente Polyphone wiedergeben.

61. Datenspeicher nach einem der Ansprüche 58 bis 60, dadurch gekennzeichnet, daß eine Gruppe von Audiosegmenten zur Verfügung gestellt wird, deren verwendete Teile mit einem statischen Phon beginnen, wobei die statischen Phone Vokale, Diphtonge, Liquide, Frikative, Vibranten und Nasale umfassen.

62. Datenspeicher nach einem der Ansprüche 58 bis 61, dadurch gekennzeichnet, daß Audiosegmente zur Verfügung gestellt werden, die geeignet sind in akustische Signale umgewandelt zu werden.

63. Datenspeicher nach einem der Ansprüche 58 bis 62, der zusätzlich Informationen enthält, um eine Variation der Frequenz, Dauer und Gesamtamplitude der verwendeten Teile einzelner Audiosegmente und deren Amplituden in verschiedenen Frequenzbereichen in Abhängigkeit von Eigenschaften der zu synthetisierenden akustischen Daten durchzuführen.

64. Datenspeicher nach einem der Ansprüche 58 bis 63, der zusätzlich Informationen enthält, die eine Variation Frequenz, Dauer und Gesamtamplitude der verwendeten Teile einzelner Audiosegmente und deren Amplituden in verschiedenen Frequenzbereichen in einem Bereich betreffen, in dem der Moment der Konkatenation liegt.

65. Datenspeicher nach einem der Ansprüche 58 bis 64, der zusätzlich verkettet Audiosegmente zur Verfügung stellt, deren Moment der Konkatenation bei einer Nullstelle der verwendeten Teile des ersten und/oder zweiten Audiosegmentes liegt.

66. Datenspeicher nach einem der Ansprüche 58 bis 65, der zusätzlich Informationen in Form von heuristischem Wissen enthält, die die Auswahl der einzelnen Audiosegmente, deren Variation sowie die Konkatenation der Audiosegmente betreffen.

67. Tonträger, der Daten enthält, die zumindest teilweise synthetisierte akustische Daten sind, die
- mit dem Verfahren nach Anspruch 1, oder
- mit der Vorrichtung nach Anspruch 16, oder
- unter Verwendung des Datenträgers nach Anspruch 45, oder
- unter Verwendung des Datenspeichers nach Anspruch 58

erzeugt wurden.

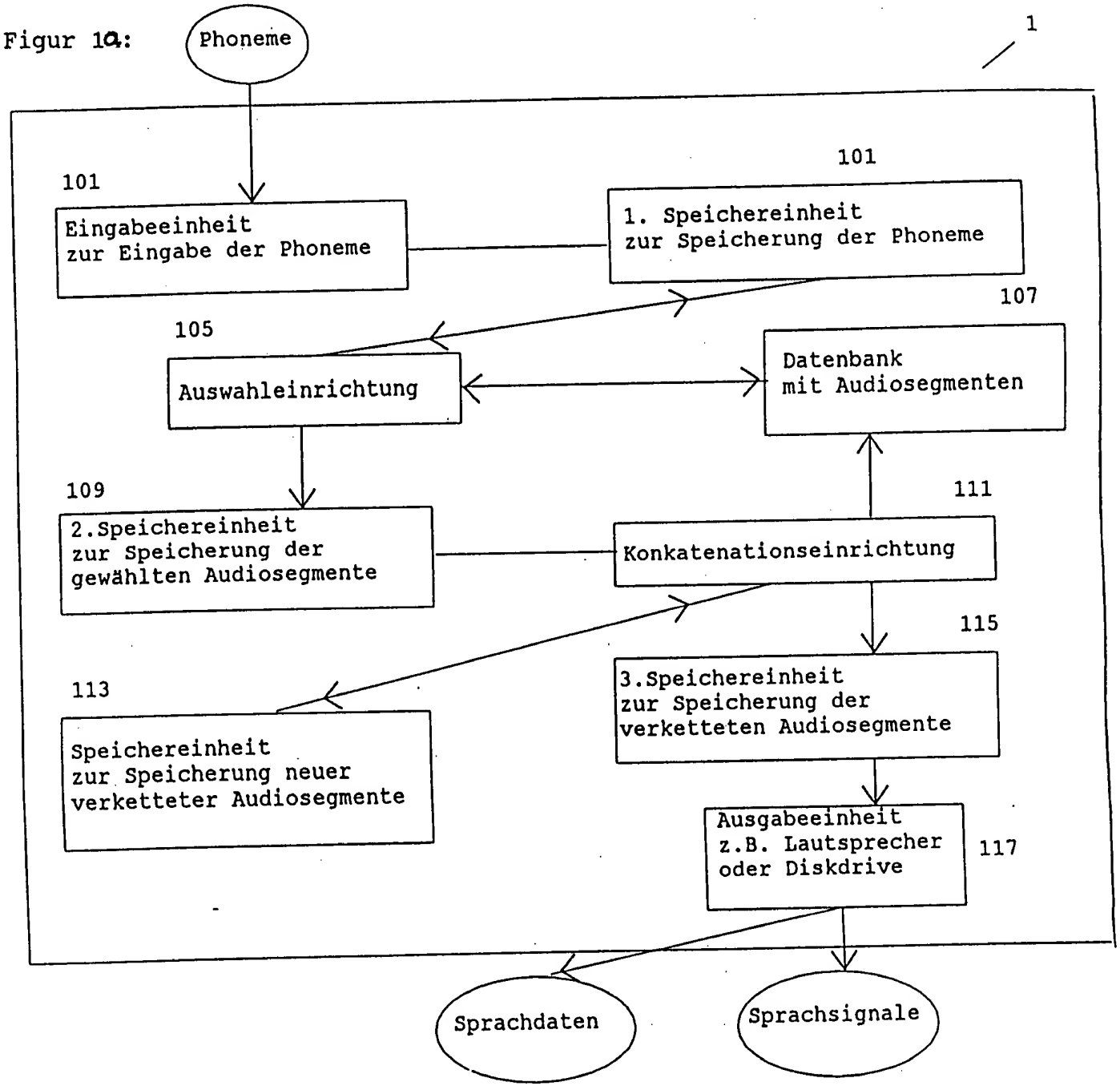
68. Tonträger nach Anspruch 67, dadurch gekennzeichnet, daß die synthetisierten akustischen Daten synthetisierte Sprachdaten sind.

69. Tonträger, der Daten enthält, die zumindest teilweise synthetisierte akustische Daten sind, die synthetisierte Sprachsignale nach Anspruch 32 sind.

Zusammenfassung

Die Erfindung ermöglicht es, beliebige akustische Daten durch eine Konkatenation einzelner Audiosegmente zu synthetisieren, wobei die Momente, zu denen die jeweilige Konkatenation zweier aufeinander folgender Audiosegmente erfolgt, in Abhängigkeit von Eigenschaften der Audiosegmente festgelegt werden. Auf diese Weise können synthetisierte akustische Daten erzeugt werden, die sich nach einer Umwandlung in akustische Signale nicht von entsprechenden natürlich erzeugen akustischen Signalen unterscheiden. Insbesondere erlaubt es die Erfindung, synthetisierte Sprachdaten unter Berücksichtigung koartikulatorischer Effekte durch Konkatenation einzelner Sprachsegmente zu erzeugen. Die so zur Verfügung gestellten Sprachdaten können in Sprachsignale umgewandelt werden, die von einer natürlich gesprochenen Sprache nicht zu unterscheiden sind.

Figur 1a:


Phoneme
1
101
Eingabeeinheit
zur Eingabe der Phoneme
101
1. Speichereinheit
zur Speicherung der Phoneme
107
105
Auswahleinrichtung
Datenbank
mit Audiosegmenten
109
2.Speichereinheit
zur Speicherung der
gewählten Audiosegmente
111
Konkatenationseinrichtung
115
3.Speichereinheit
zur Speicherung der
verketteten Audiosegmente
113
Speichereinheit
zur Speicherung neuer
verketteter Audiosegmente
Ausgabeeinheit
z.B. Lautsprecher
oder Diskdrive
117
Sprachdaten
Sprachsignale

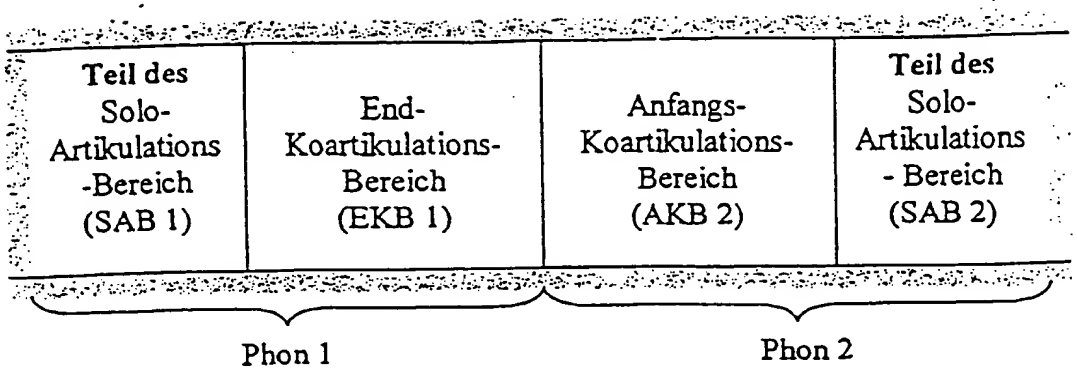
M06-09-99

Figur 1b:

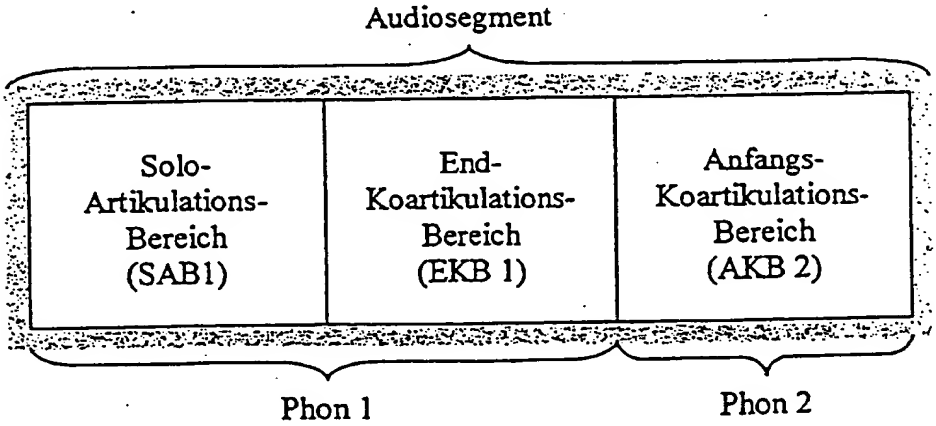
Anfangs-Koartikulations-Bereich (AKB)	Solo-Artikulations-Bereich (SAB)	End-Koartikulations-Bereich (EKB)

06.09.99

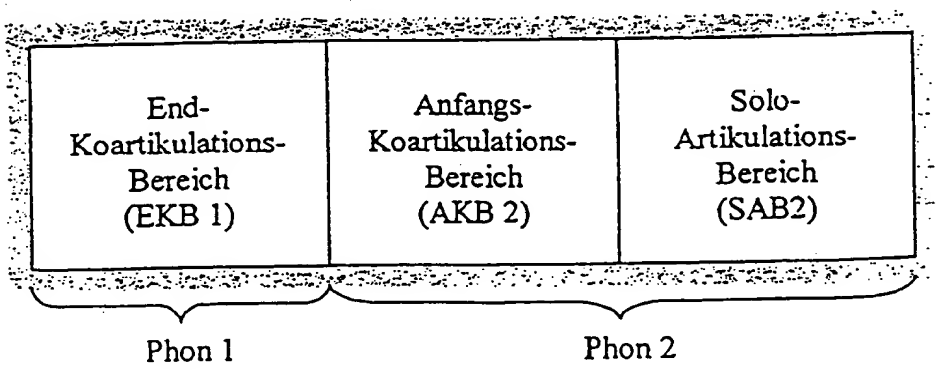
Figur 2a:



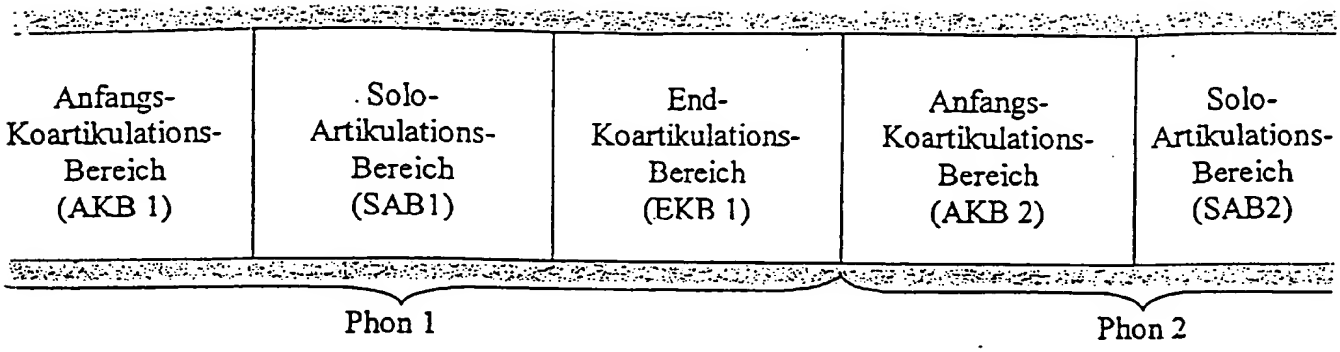
Figur 2b:



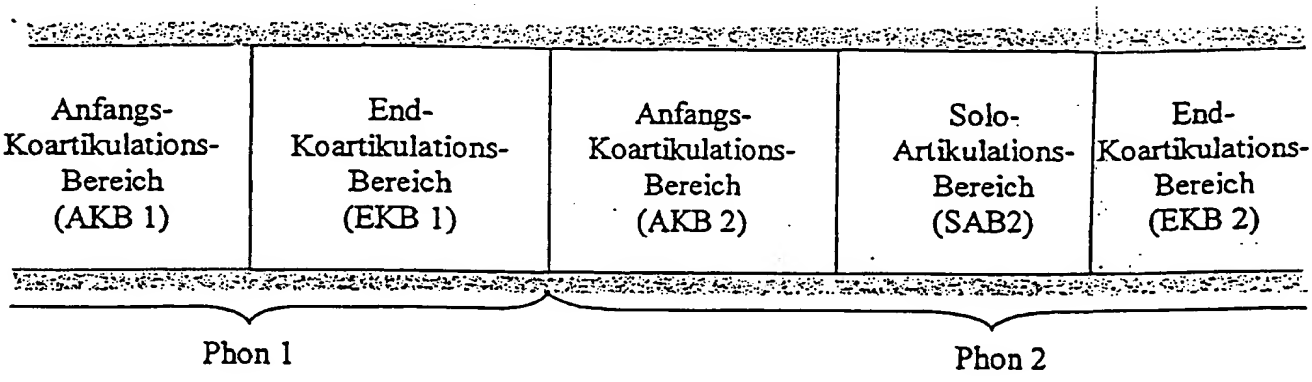
Figur 2c:



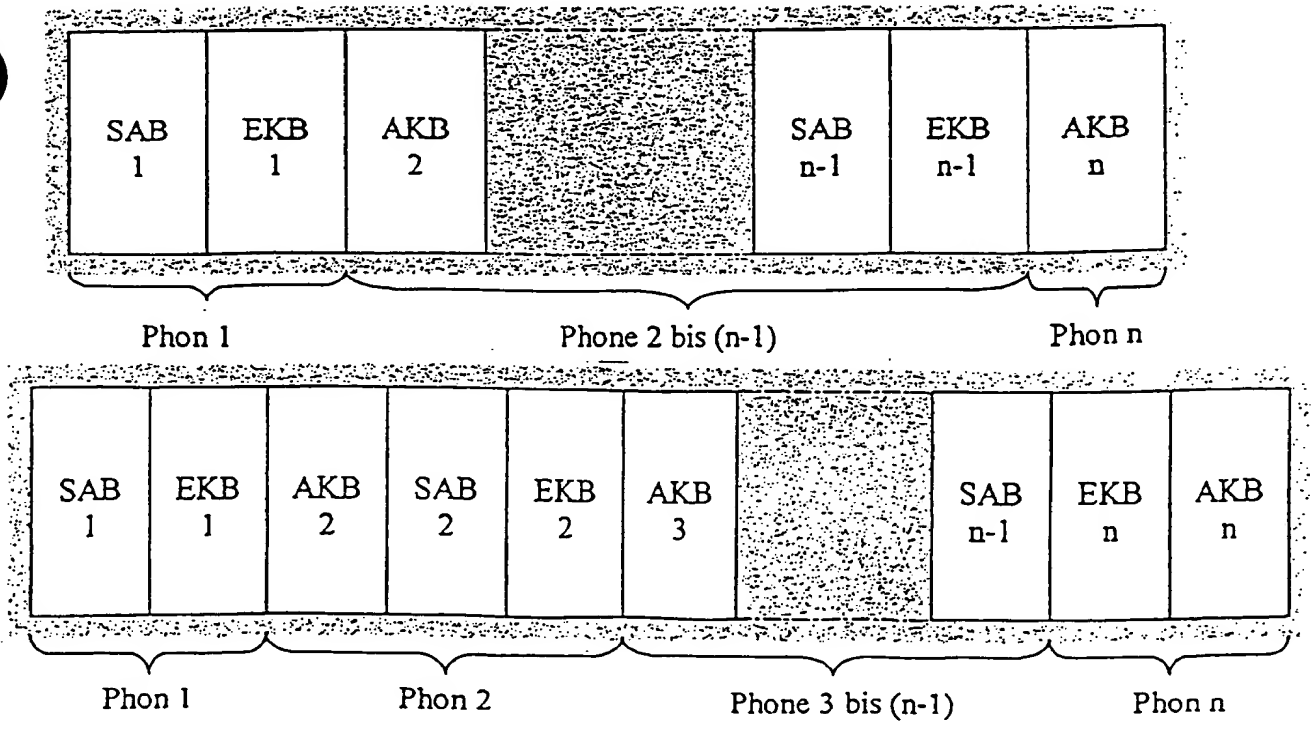
Figur 2d:



Figur 2e:

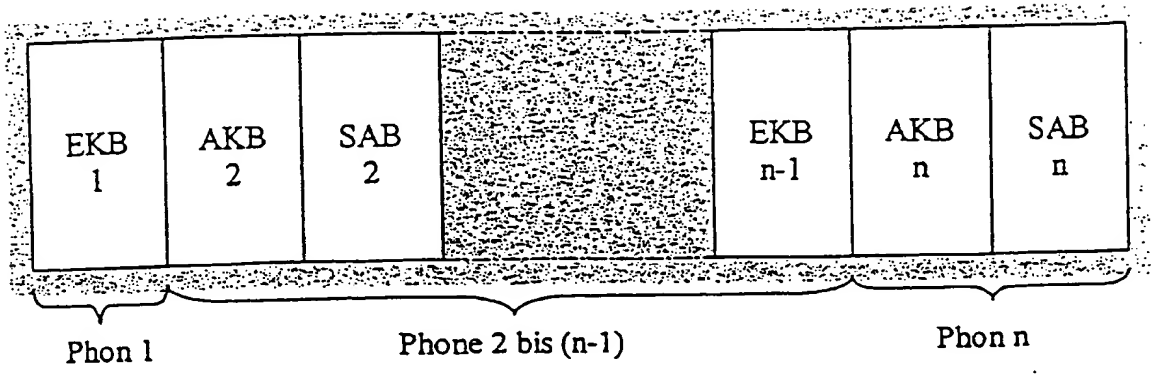


Figur 2f:

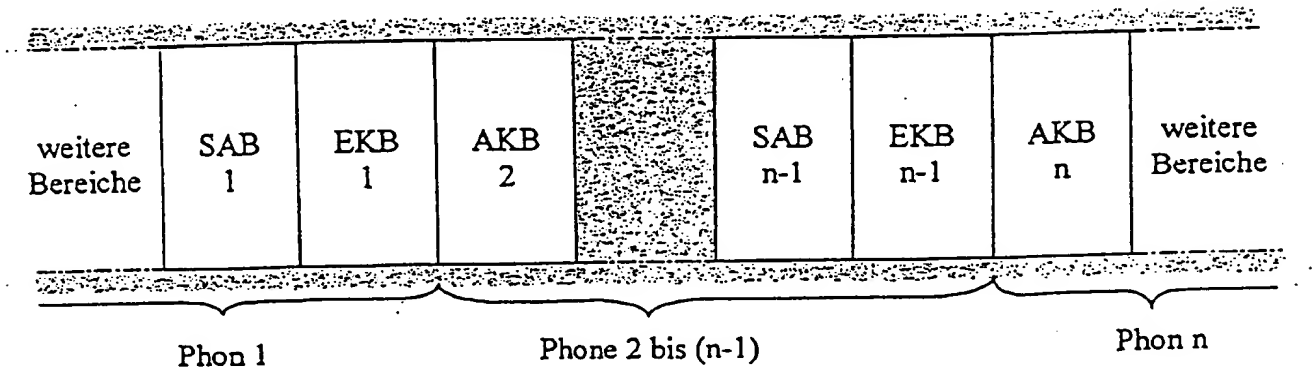


06.09.99

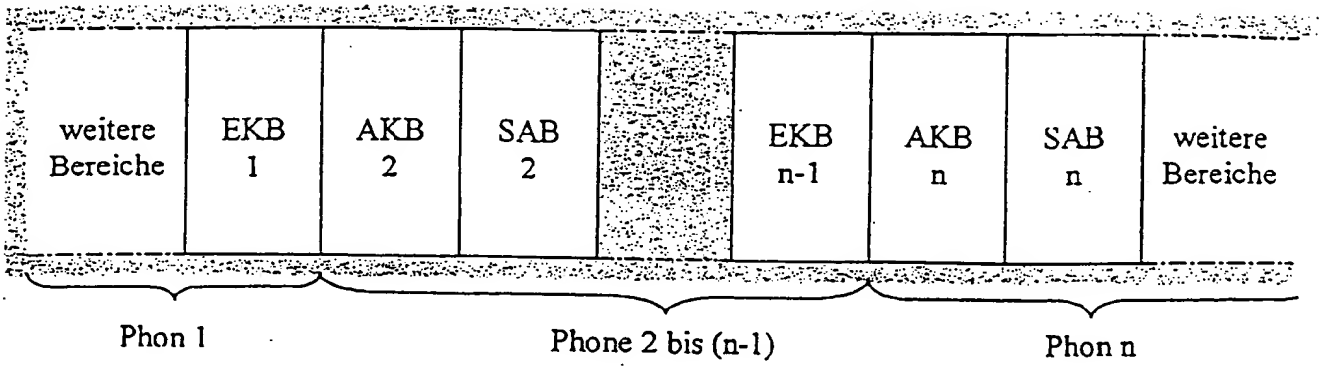
Figur 2g:



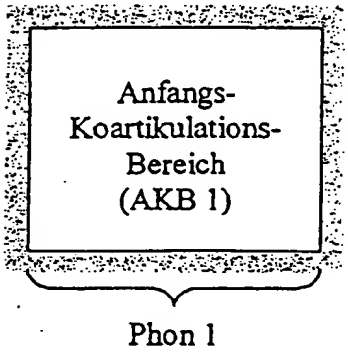
Figur 2h:

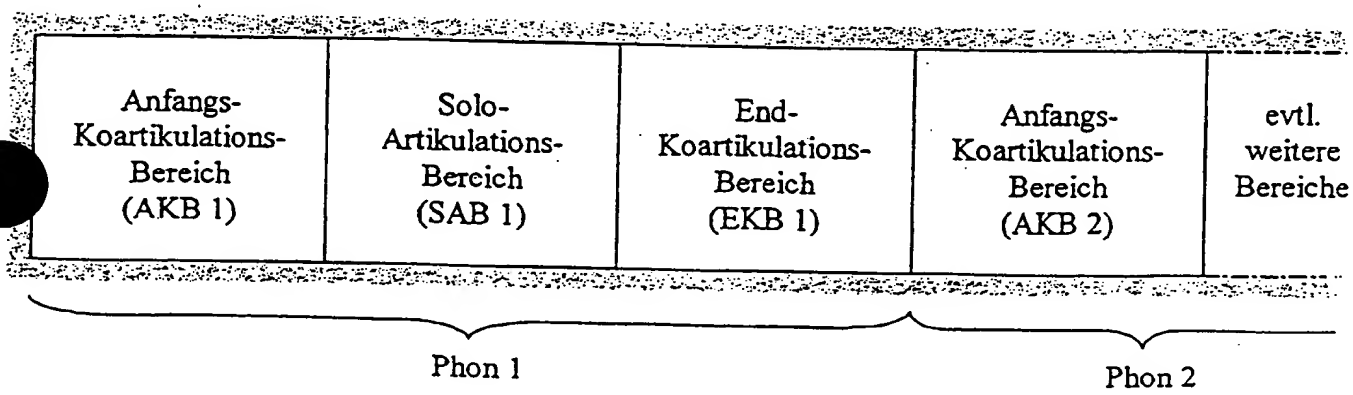


Figur 2i:

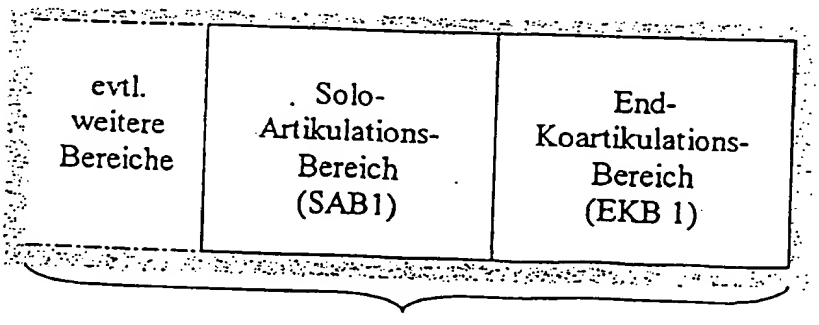


Figur 2j:

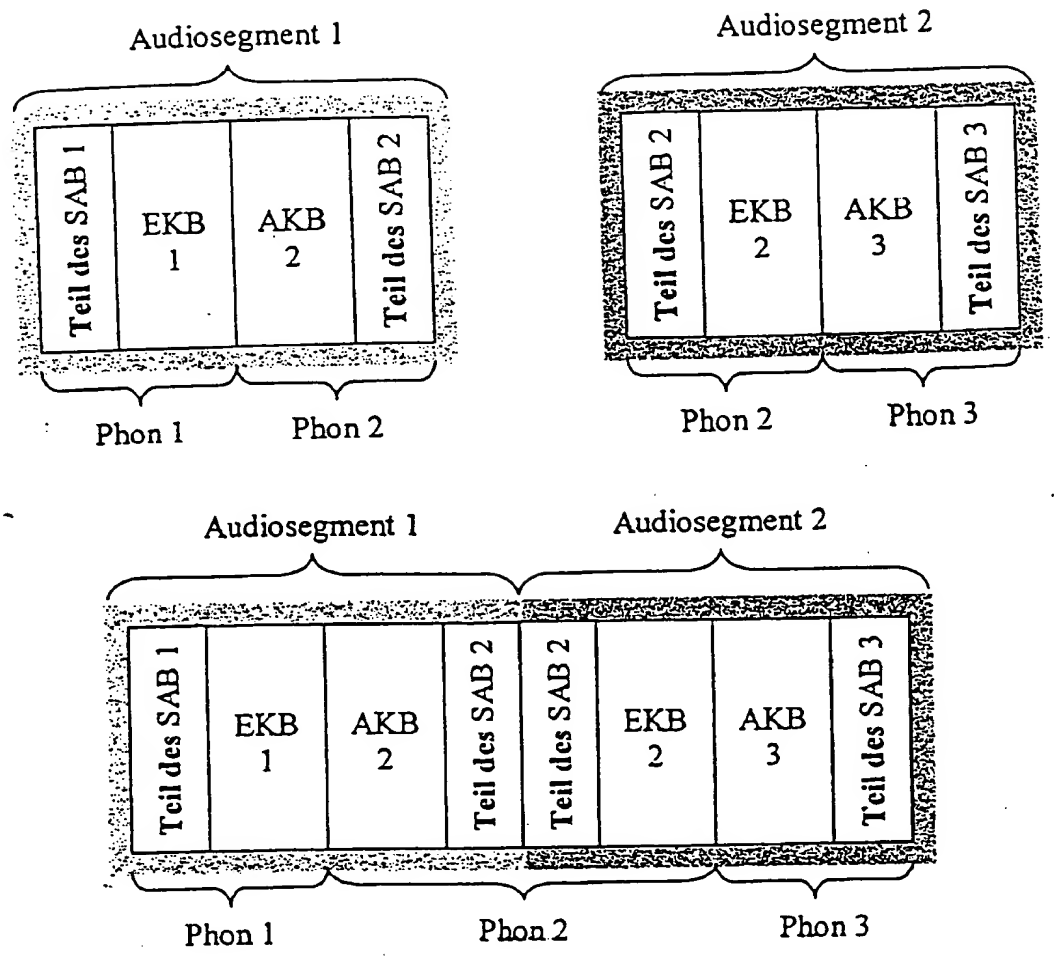




Figur 2k:

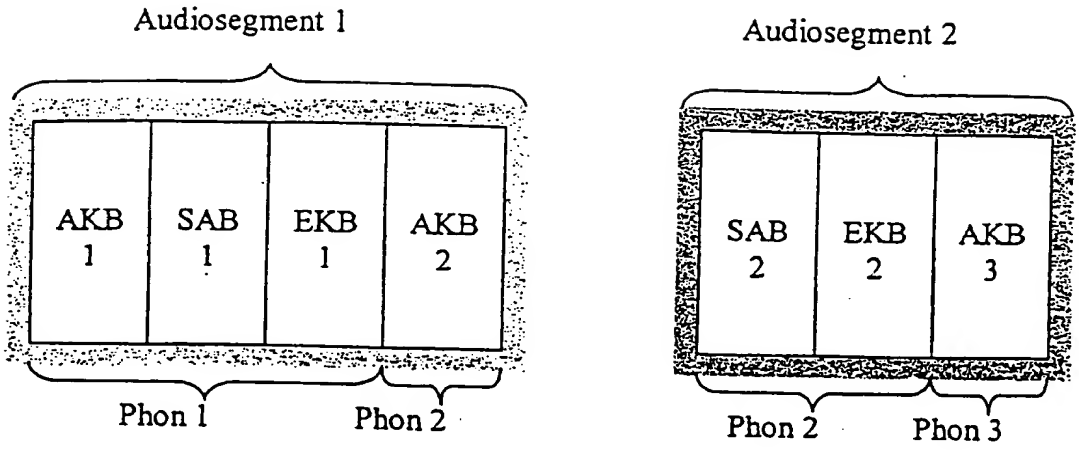


Figur 3a:

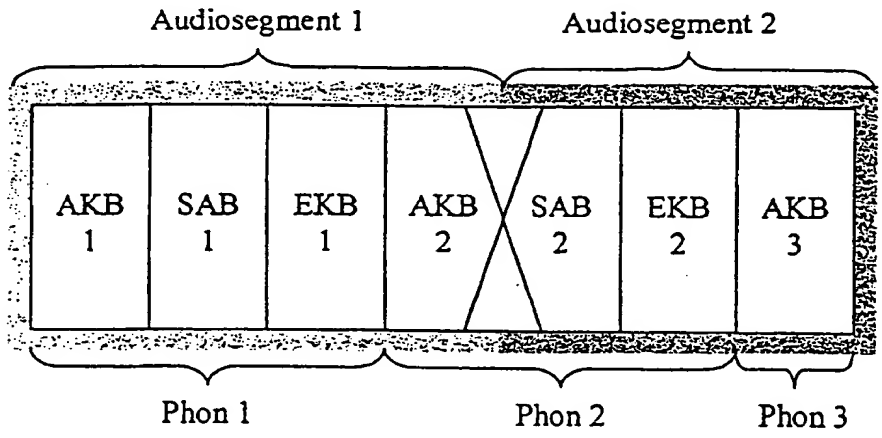

Audiosegment 1
Teil des SAB 1
EKB 1
AKB 2
Teil des SAB 2
Phon 1
Phon 2
Audiosegment 2
Teil des SAB 2
EKB 2
AKB 3
Teil des SAB 3
Phon 2
Phon 3
Audiosegment 1
Audiosegment 2
Teil des SAB 1
EKB 1
AKB 2
Teil des SAB 2
Teil des SAB 2
EKB 2
AKB 3
Teil des SAB 3
Phon 1
Phon 2
Phon 3

06.09.99

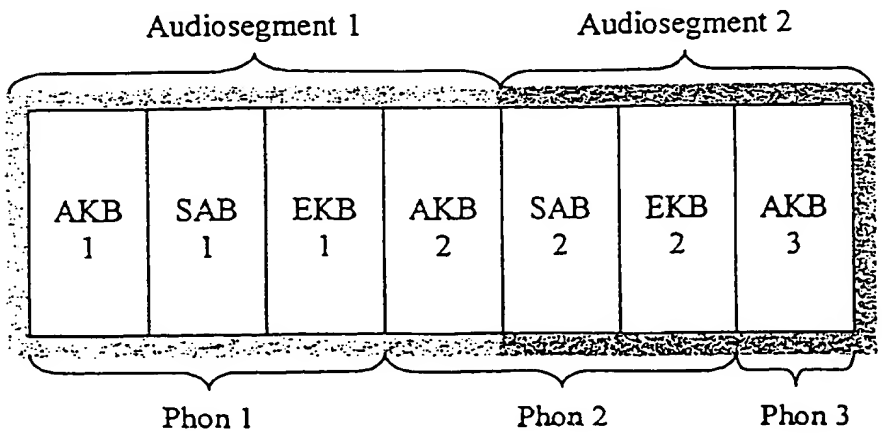
Figur 3b:



Figur 3bI:

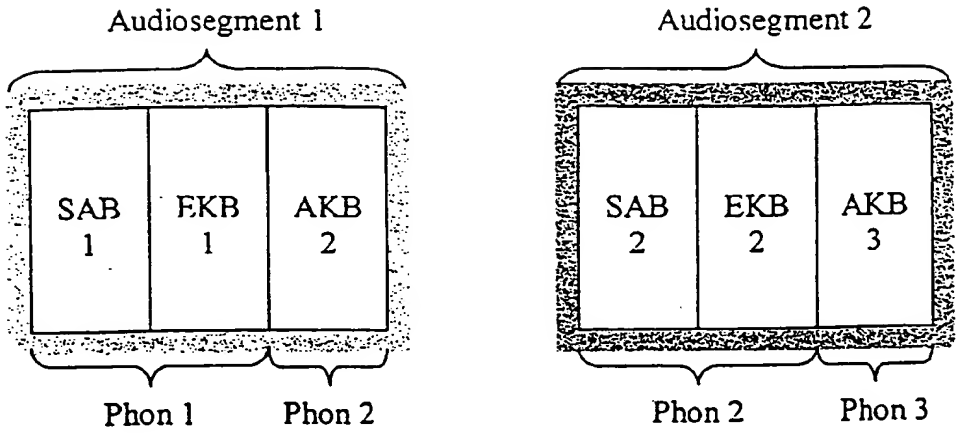


Figur3bII:

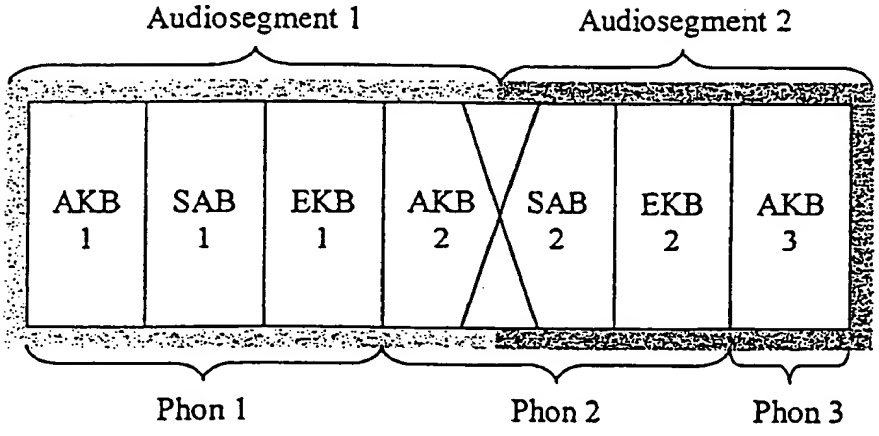


06-09-99

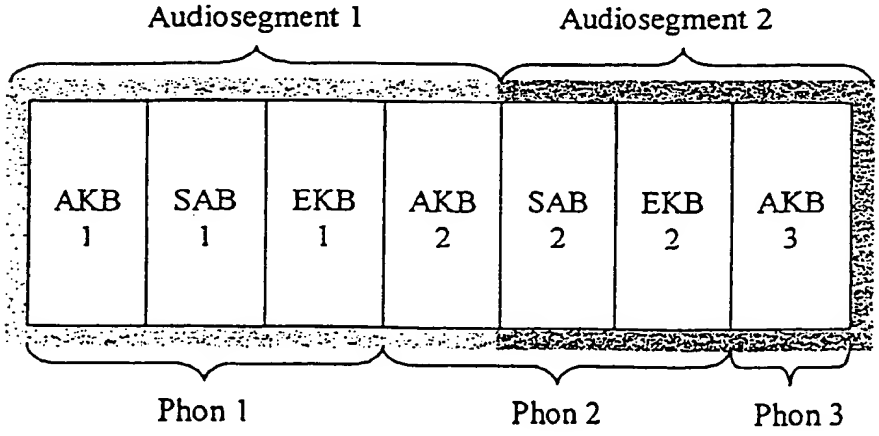
Figur 3b:



Figur 3c I:

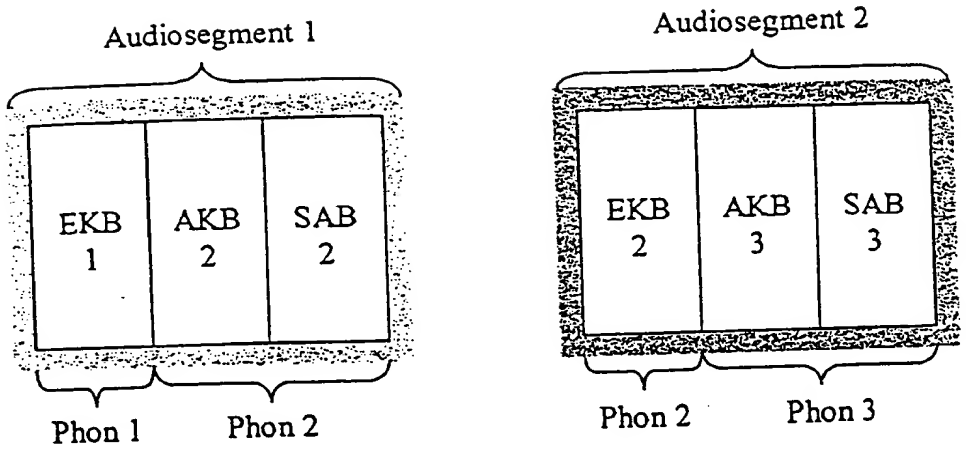


Figur 3c II:

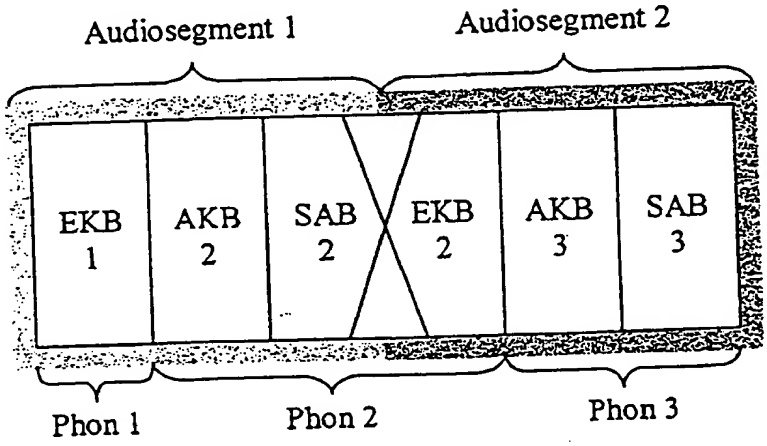


M 06.09.99

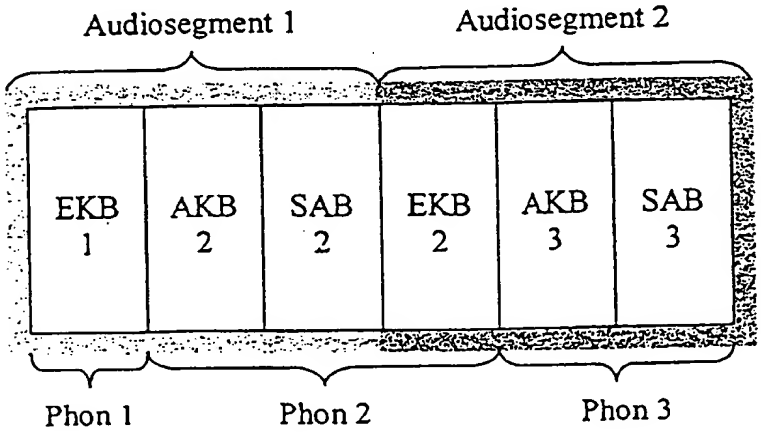
Figur 3d:



Figur 3d I:

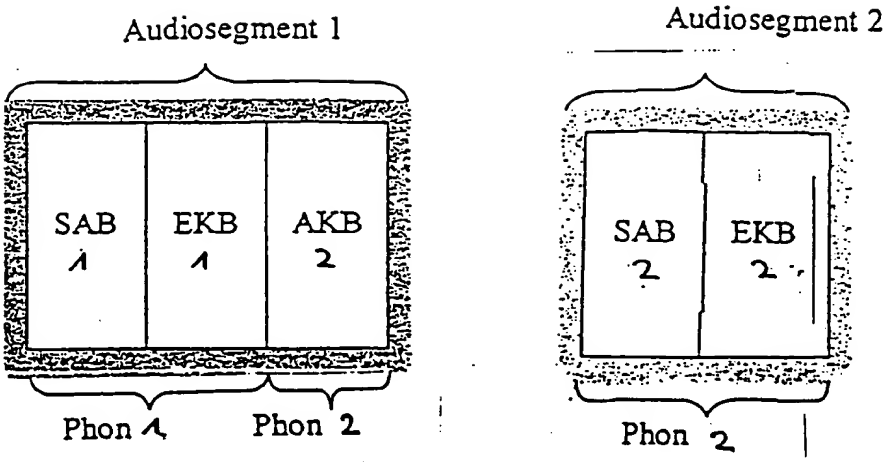


Figur 3d II:

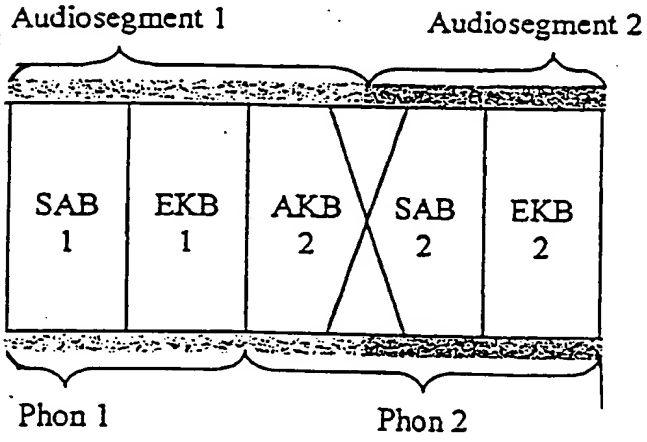


M06-09-99

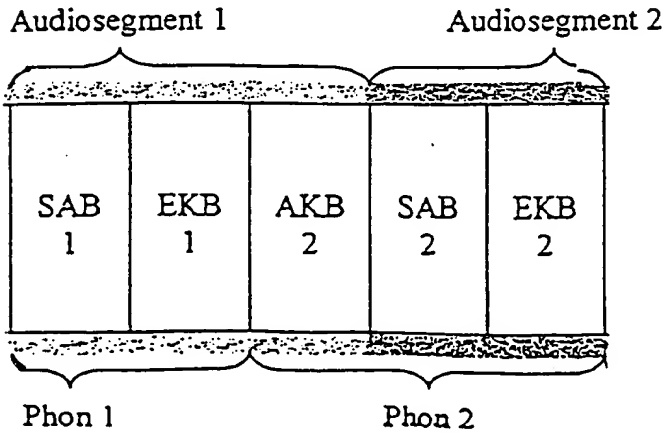
Figur 3, e:



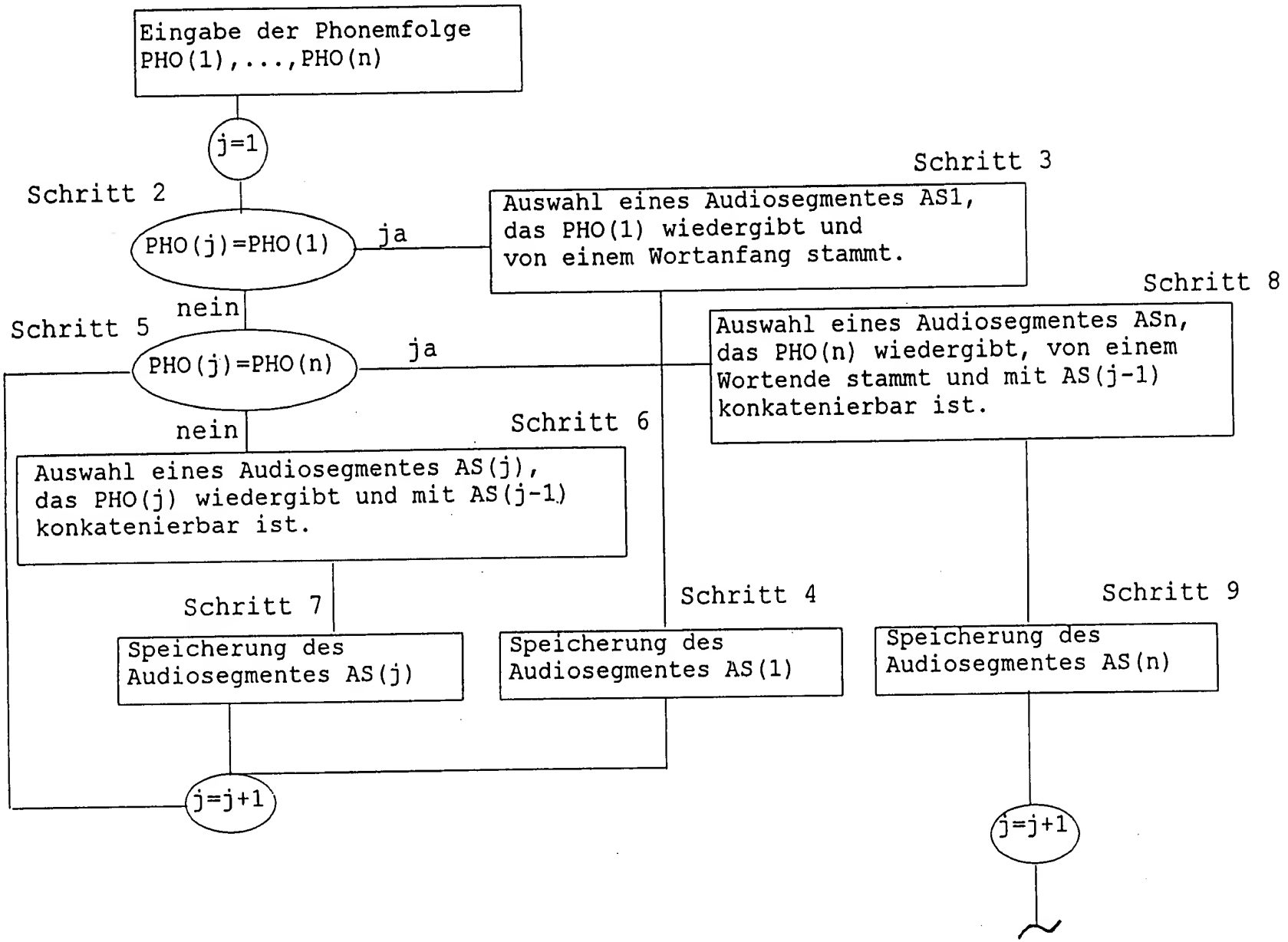
Figur 3 eI:



Figur 3 eII:



Figur 4 Teil 1



Figur 4 Teil 2